frontiers | Research Topics

Deep learning to disease prediction on next generation sequencing and biomedical imaging data

Edited by
Saurav Mallik, Junichi Iwata, Ruifeng Hu
and Tapas Si

Published in
Frontiers in Genetics

ISSN 1664-8714
ISBN 978-2-8325-3281-2
DOI 10.3389/978-2-8325-3281-2

Deep learning to disease prediction on next generation sequencing and biomedical imaging data

Topic editors

Saurav Mallik — Harvard University, United States
Junichi Iwata — University of Texas Health Science Center at Houston, United States
Ruifeng Hu — Harvard Medical School, United States
Tapas Si — University of Engineering & Management, Jaipur, Rajasthan, India

Citation

Mallik, S., Iwata, J., Hu, R., Si, T., eds. (2023). *Deep learning to disease prediction on next generation sequencing and biomedical imaging data*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3281-2

Table of contents

frontiers | Frontiers in Genetics

TYPE Editorial
PUBLISHED 04 August 2023
DOI 10.3389/fgene.2023.1260940

OPEN ACCESS

EDITED AND REVIEWED BY
Richard D. Emes,
Nottingham Trent University,
United Kingdom

*CORRESPONDENCE
Saurav Mallik,
✉ sauravmtech2@gmail.com,
✉ smallik@hsph.harvard.edu

RECEIVED 18 July 2023
ACCEPTED 26 July 2023
PUBLISHED 04 August 2023

CITATION
Mallik S, Iwata J, Hu R and Si T (2023), Editorial: Deep learning for disease prediction in next-generation sequencing and biomedical imaging data. *Front. Genet.* 14:1260940.
doi: 10.3389/fgene.2023.1260940

Editorial: Deep learning for disease prediction in next-generation sequencing and biomedical imaging data

Saurav Mallik[1]*, Junichi Iwata[2], Ruifeng Hu[3] and Tapas Si[4]

[1]Department of Environmental Health, Harvard T. H. Chan School of Public Health, Boston, MA, United States, [2]School of Dentistry, The University of Texas Health Science Center at Houston, Houston, TX, United States, [3]Department of Neurology, Brigham and Women's Hospital and Harvard Medical School Boston, Boston, MA, United States, [4]Department of Computer Science and Engineering, University of Engineering and Management, Jaipur, Gurukul, Jaipur, Rajasthan, India

KEYWORDS

deep learning, machine learning, next-generation sequencing, biomedical imaging, optimization algorithm

Editorial on the Research Topic
Deep learning to disease prediction on next-generation sequencing and biomedical imaging data

Computational learning, especially deep learning and machine learning, has had a huge impact. This Research Topic gathered articles on these two fundamental concepts which show how deep learning and machine learning approaches have been applied to array-based biomedical data such as next-generation sequencing (NGS) and medical imaging data.

Overall, our Research Topic published 11 articles of which 8 covered research on array-based data, while the remaining 3 articles belonged to studies on biomedical imaging. Among them, She et al. propose a joint mathematical model integrating a random forest classifier and artificial neural network (ANN) for the possible diagnosis of the estrogen-dependent inflammatory disease endometriosis. The method utilizes publicly available gene expression datasets in the Gene Expression Omnibus (GEO) and estimated seven significant differentially expressed genes (DEGs) (viz., COMT, NAA16, CCDC22, EIF3E, AHI1, DMXL2, and CISD3) through the random forest classifier, while three of them (AHI1, DMXL2, and CISD3) were novel signatures useful for the pathogenesis of endometriosis. Related KEGG pathway and GENE Ontology analysis is also performed to obtain the biological significance of the signatures. Niu et al. conduct a comprehensive bioinformatic analysis to determine the potential diagnostic and prognostic genetic markers for gastric cancer. In this study, several markers (COL1A1, COL5A2, P4HA3, and SPARC) yielded high scores in the prognosis and diagnosis of gastric cancer, hence they are named as the respective diagnosis and prognosis markers for the disease. A second study conducted by the same team Niu et al. focuses on an extracellular matrix protein, prolyl 4-hydroxylase subunit alpha 3 (P4HA3), and thus performed an extensive protein-protein interaction and prognosis analysis in terms of correlating it with immune infiltration in the gastric Cancer. Another study was conducted by Gu et al. in which an angiogenic factor-based gene signature is identified that had a significant response in patients' survival, disease prognosis, and immunotherapy in non-small-cell lung cancer, a common malignancy. The corresponding model had good discrimination and calibration and may predict the disease prognosis of treatment in the respective clinical practice. Wei et al. provide a comprehensive bioinformatic analysis to determine a potential prognostic genetic marker (viz., GNG7) for the lung adenocarcinoma that correlates with the immune infiltrates. Wen et al. introduce a framework by integrating several machine learning algorithms to determine whether hub genes are useful for the diagnosis of ankylosing spondylitis by validating several respective gene expression datasets. A novel machine learning and optimization framework termed as 3-factor penalized non-negative matrix factorization-based multiple kernel learning with the soft margin hinge loss (3PNMF-MKL) is proposed by Mallik et al. where two consecutive steps, namely, multi-modal data integration and gene signature discovery are conducted.

Essential genes are required for critical cellular activities in the overall survival of many species. Rout et al. conduct an extensive analysis to determine the discriminant features (genes) from the stationary pattern of the nucleotide bases (A, T, G, C) and their respective application towards the classification of the essential gene.

From the imaging point of view, a dual-input convolution neural network (CNN) with the local interpretable model-agnostic explanation (LIME) and Shapley additive explanation (SHAP) is utilized to predict the discrete subtypes of brain tumors, viz., glioma, meningioma, and pituitary through the Magnetic Resonance Imaging (MRI) of brain (Gaur et al.). Another study was conducted by Sharma et al. where the likelihood of a colorectal cancer patient dying could be significantly decreased through the early diagnosis as well as treatment of the pre-cancerous polyps. Sharma et al. develop an ensemble-based deep CNN model that helped to identify the polyps from a colonoscopy video with a higher accuracy which outperformed the existing methodologies (viz., ResNet101, Xception, and GoogleNet). The projections of the lateral chest radiograph (chest X-rays or, CXR) of children with clinically suspected pulmonary tuberculosis (TB) yielded a significant enhancement in the overall sensitivity of the enlarged lymph nodes. A model-level ensemble was built through the fine-tuned CNN and Vision Transformers (ViT) models by Rajaraman et al. to detect the TB-consistent outcomes in the lateral CXRs, and finally, a significantly better classification performance could be obtained.

This Research Topic covers articles on developing frameworks/tools/algorithms for handling next-generation sequencing (NGS) array-based data as well as medical imaging data. It is expected that future machine/deep learning software will be increasingly helpful for biomedical and healthcare researchers to realize the utilization of machine/deep learning and optimization to improve the overall research quality and integrity in disease diagnosis and potential therapeutic use.

Author contributions

SM: Conceptualization, Investigation, Methodology, Software, Supervision, Validation, Writing–original draft, Writing–review and editing. JI: Data curation, Formal Analysis, Project administration, Software, Writing–original draft. RH: Data curation, Investigation, Methodology, Validation, Writing–original draft. TS: Conceptualization, Data curation, Software, Visualization, Writing–original draft.

Acknowledgments

We are highly grateful to the authors for source of all articles regarding this Research Topic.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 February 2022
doi: 10.3389/fgene.2022.864724

Detecting Tuberculosis-Consistent Findings in Lateral Chest X-Rays Using an Ensemble of CNNs and Vision Transformers

Sivaramakrishnan Rajaraman[1]*, *Ghada Zamzmi*[1], *Les R. Folio*[2] *and Sameer Antani*[1]

[1]Computational Health Research Branch, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States, [2]Moffitt Cancer Center, Tampa, FL, United States

Research on detecting Tuberculosis (TB) findings on chest radiographs (or Chest X-rays: CXR) using convolutional neural networks (CNNs) has demonstrated superior performance due to the emergence of publicly available, large-scale datasets with expert annotations and availability of scalable computational resources. However, these studies use only the frontal CXR projections, i.e., the posterior-anterior (PA), and the anterior-posterior (AP) views for analysis and decision-making. Lateral CXRs which are heretofore not studied help detect clinically suspected pulmonary TB, particularly in children. Further, Vision Transformers (ViTs) with built-in self-attention mechanisms have recently emerged as a viable alternative to the traditional CNNs. Although ViTs demonstrated notable performance in several medical image analysis tasks, potential limitations exist in terms of performance and computational efficiency, between the CNN and ViT models, necessitating a comprehensive analysis to select appropriate models for the problem under study. This study aims to detect TB-consistent findings in lateral CXRs by constructing an ensemble of the CNN and ViT models. Several models are trained on lateral CXR data extracted from two large public collections to transfer modality-specific knowledge and fine-tune them for detecting findings consistent with TB. We observed that the weighted averaging ensemble of the predictions of CNN and ViT models using the optimal weights computed with the Sequential Least-Squares Quadratic Programming method delivered significantly superior performance (MCC: 0.8136, 95% confidence intervals (CI): 0.7394, 0.8878, p < 0.05) compared to the individual models and other ensembles. We also interpreted the decisions of CNN and ViT models using class-selective relevance maps and attention maps, respectively, and combined them to highlight the discriminative image regions contributing to the final output. We observed that (i) the model accuracy is not related to disease region of interest (ROI) localization and (ii) the bitwise-AND of the heatmaps of the top-2-performing models delivered significantly superior ROI localization performance in terms of mean average precision [mAP@(0.1 0.6) = 0.1820, 95% CI: 0.0771,0.2869, p < 0.05], compared to other individual models and ensembles. The code is available at https://github.com/sivaramakrishnan-rajaraman/Ensemble-of-CNN-and-ViT-for-TB-detection-in-lateral-CXR.

Keywords: chest radiographs, CNN, deep learning, tuberculosis classification and localization, vision transformers, ensemble learning, significance analysis

OPEN ACCESS

Edited by:
Tapas Si,
Bankura Unnayani Institute of Engineering, India

Reviewed by:
Utpal Nandi,
Vidyasagar University, India
Wellington Pinheiro dos Santos,
Federal University of Pernambuco, Brazil

****Correspondence:***
Sivaramakrishnan Rajaraman
sivaramakrishnan.rajaraman@nih.gov

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *28 January 2022*
Accepted: *10 February 2022*
Published: *24 February 2022*

Citation:
Rajaraman S, Zamzmi G, Folio LR and Antani S (2022) Detecting Tuberculosis-Consistent Findings in Lateral Chest X-Rays Using an Ensemble of CNNs and Vision Transformers. Front. Genet. 13:864724. doi: 10.3389/fgene.2022.864724

1 INTRODUCTION

Artificial intelligence (AI) methods, particularly deep learning (DL)-based convolutional neural network (CNN) models, have demonstrated remarkable performance in natural and medical computer vision applications (Schmidhuber, 2015). Considering chest-X-ray (CXR) analysis, CNN models have outperformed conventional machine learning (ML) methods for semantic segmentation, classification, and object detection, among other tasks (Wang et al., 2017; Irvin et al., 2019; Bustos et al., 2020).

Research on detecting Tuberculosis (TB)-consistent findings in CXRs using DL methods has demonstrated superior performance due to the emergence of publicly available, large-scale datasets with expert annotations and availability of scalable computational resources (Jaeger et al., 2014; Lakhani and Sundaram, 2017; Sivaramakrishnan et al., 2018; Pasa et al., 2019; Rajaraman and Antani, 2020). However, these studies only use the frontal CXR projections, i.e., the posterior-anterior (PA), and the anterior-posterior (AP) views, for analysis and decision-making. To the best of our knowledge, lateral CXR projections have, heretofore, not been used for AI detection approaches to pulmonary diseases before this work. Lateral CXR projections of children with clinically suspected pulmonary TB, in addition to the conventional frontal projections, are critical and showed an increase in the detection sensitivity of enlarged lymph nodes by 1.8% and specificity by 2.5% (Swingler et al., 2005). Further, the World Health Organization (WHO) recommends the use of lateral CXR projections to identify mediastinal or hilar lymphadenopathy (World Health Organization, 2016), especially in younger children with primary TB where a bacteriological confirmation might be challenging. As discussed in (Gaber et al., 2005), lateral CXRs provide useful spatial diagnostic information on the thoracic cage, pleura, lungs, pericardium, heart, mediastinum, and upper abdomen and help identify lymphadenopathy in children with primary TB (Gaber et al., 2005). Another study (Herrera Diaz et al., 2020) discusses the current national Canadian guidelines suggesting using lateral CXR projections for TB screening upon admission to long-term care facilities. These studies underscore the importance of using lateral CXR projections as they carry useful information on disease manifestation and progression; hence, this study aims to explore these least studied types of CXR projection (the lateral) and propose a novel approach for detecting TB-consistent findings.

Recently, Vision Transformers (ViTs) (Zhai et al., 2021) with built-in self-attention mechanisms have demonstrated comparable performance to CNNs in natural and medical visual recognition tasks, while requiring fewer computational resources. Several studies (Liu and Yin, 2021; Shome et al., 2021; Park et al., 2022) used ViTs to improve pulmonary disease detection in frontal CXRs to detect manifestations consistent with COVID-19 disease. Another study (Duong et al., 2021) used a ViT model to detect TB-consistent findings in frontal CXRs and obtained an accuracy of 97.72%. The promising performance of ViT models in medical visual recognition tasks is constrained by sparse data availability (Zhai et al., 2021). Unlike CNN models, ViT models lack intrinsic biases, i.e., the properties of translation equivariance, which is the similarity in processing different image parts regardless of their absolute position, and they do not consider the relationship between the neighboring image pixels. Further, the computational complexity of ViT models increases with the input image resolution resulting in demand for a higher resource. In contrast, CNN models have shown promising performance even with limited data due to their inherent inductive bias characteristics that help in convergence and generalization. However, CNN models do not encode the relative position of different image features and may require large receptive fields to encode the combination of these features and capture long-range dependencies in an input image. This leads to increased convolutional kernel sizes and subsequently the computational complexity (Alzubaidi et al., 2021). A potential solution could be to exploit the advantages of both models, i.e., CNNs and ViTs toward decision-making for the task under study.

Several ensemble methods including majority voting, averaging, weighted averaging, and stacking, have been studied for medical visual recognition tasks (Dietterich, 2000). Considering CXR analysis, particularly TB detection, ensemble methods have been widely used to improve performance in semantic segmentation, classification, and object detection tasks (Hogeweg et al., 2010; Ding et al., 2017; Islam et al., 2017; Rajaraman et al., 2018a; Rajaraman and Antani, 2020). However, to the best of our knowledge, we are not aware of studies that perform an ensemble of ViTs or an ensemble of both CNN and ViT models for disease detection, particularly detecting TB-consistent findings using lateral CXRs. The main contribution of this work is a systematic approach that benefits from constructing ensembles of the best models from both worlds (i.e., CNNs and ViTs) to detect TB-consistent findings using lateral CXRs through reduced prediction variance and improved performance.

The steps in this systematic study can be summarized as follows: (i) First, ImageNet-pretrained CNN models, viz, VGG-16 (Simonyan and Zisserman, 2015), DenseNet-121 (Huang et al., 2017), and EfficientNet-V2-B0 (Tan and Le, 2021) and the ImageNet-pretrained ViT models, viz, ViT-B/16, ViT-B/32, ViT-L/16, and ViT-L/32 (Zhai et al., 2021) are retrained on a combined selection of publicly available lateral CXR collections (Rajpurkar et al., 2017; Bustos et al., 2020). This step is performed to convert the weight layers specific to the lateral CXR modality and learn to classify normal and abnormal lateral CXRs; (ii) Next, the retrained models are used to transfer the lateral CXR modality-specific knowledge to improve performance in the related task of classifying lateral CXRs as showing no abnormalities or other findings that are consistent with TB; (iii)The predictions of the top-K (K = 2, 3, 5, 7) models are combined using several ensemble methods such as majority voting, simple averaging, and weighted averaging using the optimal weights derived with the Sequential Least-Squares Quadratic Programming (SLSQP) algorithm (Gupta and Gupta, 2018). We construct a “model-level” ensemble of the CNN and ViT models by flattening, concatenating the features from their deepest layers, and adding the classification layers to

TABLE 1 | Datasets and their respective patient-level train/test splits. Data in parenthesis denotes the 90/10 train/test splits. A part of the lateral CXRs in the PadChest CXR collection that show no abnormalities and those with TB-consistent manifestations are used for fine-tuning. The rest of the data from the PadChest and CheXpert lateral CXR collections are used for CXR modality-specific pretraining.

Dataset	CXR modality-specific pretraining		Fine-tuning	
	Abnormal	Normal	TB	Normal
PadChest	32923 (29631/3292)	13698 (12328/1370)	530 (477/53)	530 (477/53)
CheXpert	23633 (21270/2363)	4717 (4245/472)	-	-

classify the lateral CXRs to their respective categories; (iv) We also interpret CNN and ViT model decisions through the use of class-selective relevance maps (CRM) (Kim et al., 2019) and attention maps, respectively, and construct an ensemble of these heatmaps and attention maps using several ensemble methods. Finally, we analyze and report statistical significance in the results obtained using the individual models and their ensembles using confidence intervals (CIs) and p values.

2 MATERIALS AND METHODS

2.1 Datasets

The following publicly available datasets are used in this study:

CheXpert CXR dataset: The authors in (Irvin et al., 2019) released a collection of frontal and lateral CXR projections, showing normal lungs, and other pulmonary abnormalities. The dataset contains 224,316 CXRs collected from 65,240 patients at the Stanford University Hospital in California. The CXRs are labeled using a natural language processing (NLP)-based automatic labeler for the presence of 14 thoracic abnormalities mentioned in radiological reports. The collection includes 23,633 lateral CXRs manifesting various pulmonary abnormalities and 4,717 lateral CXRs showing no abnormalities. In this study, the lateral CXR projections are split at the patient level into 90/10 proportions for the train and test sets and are used during CXR modality-specific pretraining.

PadChest CXR dataset: A collection of 160,000 frontal and lateral CXRs and their associated radiological reports are released by (Bustos et al., 2020). The collection includes normal and abnormal CXRs collected from 67,000 patients at the San Juan Hospital in Spain. The CXR images are automatically labeled for 174 radiographic findings, based on the Unified Medical Language System (UMLS) terminology. The collection includes 33,454 lateral CXRs manifesting several pulmonary abnormalities and 14,229 lateral CXRs showing no abnormalities. The abnormal lateral CXR collection also includes 530 CXRs collected from patients diagnosed with TB. The set of CXRs manifesting TB-consistent findings and an equal number of lateral CXRs with no abnormalities are used during the fine-tuning. The ground truth annotations for the hold-out test set consisting of 53 images, and showing findings that are consistent with TB, are provided by an expert radiologist (with >30 years of experience). The radiologist used the web-based VGG Image Annotator tool (VIA, Oxford, England) (Dutta and Zisserman, 2019) to annotate the test collection by manually setting boundary boxes for what is believed to be TB-consistent findings. **Table 1** shows the datasets, the numbers of images, and their respective patient-level train/test splits used in this study. The lateral CXR images from the PadChest and CheXpert collections are resized to 224 × 224 pixel dimensions to reduce computational overhead.

2.2 Classification Models

The following CNN and ViT Models are used in this study: (i) VGG-16 (Simonyan and Zisserman, 2015); (ii) DenseNet-121 (Huang et al., 2017); (iii) EfficientNet-V2-B0 (Tan and Le, 2021); (iv) ViT-Base (B)/16 (Zhai et al., 2021); (v) ViT-B/32 (Zhai et al., 2021); (vi) ViT-Large (L)/16 (Zhai et al., 2021); and (vii) ViT-L/32 (Zhai et al., 2021). The CNN models are selected based on their superior performance in CXR-based visual recognition tasks (Wang et al., 2017; Rajaraman et al., 2018b; Irvin et al., 2019; Rajaraman et al., 2020a). The numbers 16 and 32 in the ViT models denote the size of input image patches. The length of the input image patch sequence is inversely proportional to the square of the patch size. Thus, the ViT models with smaller patch sizes are computationally more expensive (Zhai et al., 2021). Interested readers are referred to (Wang et al., 2017; Rajaraman et al., 2018b; Irvin et al., 2019; Rajaraman et al., 2020a; Zhai et al., 2021) for a detailed description of these models' architecture.

2.3 CXR Modality-Specific Pretraining, Fine-Tuning, and Ensemble Learning

During CXR modality-specific pretraining, the CNN models are instantiated with their ImageNet pretrained weights, truncated at their optimal intermediate layers (Rajaraman et al., 2020b), and appended with the following layers: (i) a zero-padding (ZP) layer, (ii) a convolutional layer with 512 filters, each of size 3 × 3, (iii) a global averaging pooling (GAP) layer; and (iv) a final dense layer with two nodes and Softmax activation. The optimal intermediate layers are identified from pilot analyses for the task under study. The ViT models are instantiated with their pretrained weights learned from a combined selection of ImageNet and Imagenet21K datasets. These models are then truncated at the output classification token layer and appended with a flattening layer and a final dense layer with two nodes to output prediction probabilities. **Figure 1** shows the block diagram of models used in CXR modality-specific pretraining and fine-tuning stages.

The CNN and ViT models are then retrained on a combined selection of lateral CXRs from the CheXpert and PadChest datasets (**Table 1**). This process is called CXR modality-specific pretraining and it is performed to impart CXR

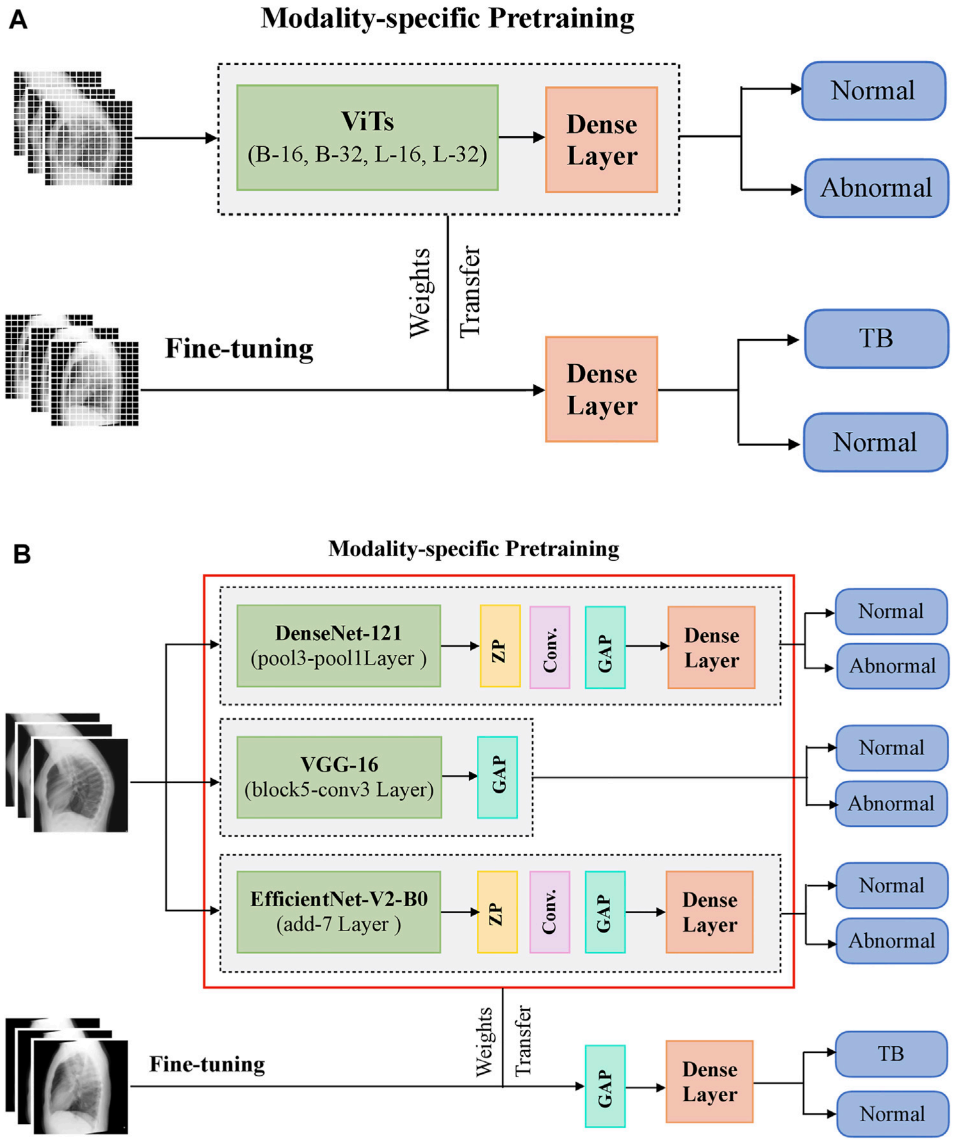


FIGURE 1 | A systematic approach of training the models during CXR modality-specific pretraining and fine-tuning stages. **(A)** ViTs and **(B)** CNNs.

modality-specific knowledge to (i) coarsely learn the characteristics of normal and abnormal lateral CXRs and (ii) convert the weight layers learned from natural images to the input CXR modality. The modality-specific pretrained CNN and ViT models are then fine-tuned to classify the lateral CXRs as showing no abnormalities or other findings that are consistent with TB. The datasets are split at the patient level into 90% for training and 10% for testing during the CXR modality-specific pretraining and finetuning stages as shown in **Table 1**. We allocated 10% of the training data for validation with a fixed seed. The training data is augmented using affine transformations such as rotation (−5, +5), horizontal flipping, width, and height shifting (−5, +5), and normalized so the image pixel values lie in the range (0, 1). During CXR modality-specific pretraining, the CNN and ViT models are trained for 100 epochs, using a stochastic gradient descent (SGD) optimizer with an initial learning rate of 1e-2 and momentum of 0.9, to minimize the categorical cross-entropy loss. We used callbacks to store model checkpoints and reduced the learning rate whenever the validation loss ceased to decrease. The best-performing model, delivering the least validation loss at the end of the training epochs is stored to predict the hold-out test set. During fine-tuning, the CXR modality-specific pretrained models are finetuned using the SGD optimizer with an initial learning rate of 1e-4 and momentum of 0.9. We used callbacks for early stopping and learning rate reduction. The best-performing model, delivering the least validation loss at the end of the training epochs is stored to predict the hold-out test set.

The top-K (K = 2, 3, 5, 7) fine-tuned models that deliver superior performance with the hold-out test set are used to construct ensembles. We constructed "prediction-level" and

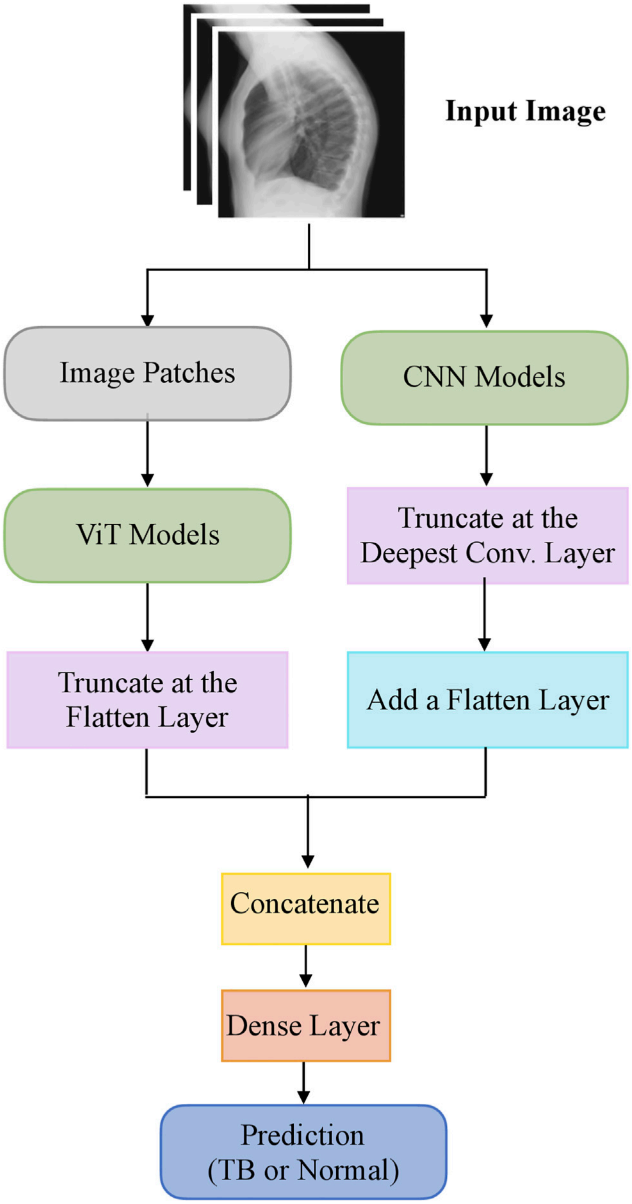

Input Image
Image Patches
CNN Models
ViT Models
Truncate at the Deepest Conv. Layer
Truncate at the Flatten Layer
Add a Flatten Layer
Concatenate
Dense Layer
Prediction (TB or Normal)

FIGURE 2 | A model-level ensemble constructed using fine-tuned CNN and ViT models.

"model-level" ensembles. At the prediction level, we used several ensemble strategies such as majority voting, simple averaging, and SLSQP-based weighted averaging to combine the top-K model predictions. For SLSQP-based weighted averaging, we computed the optimal weights by minimizing the total logarithmic loss using the SLSQP algorithm (Gupta and Gupta, 2018) to help convergence. For the model-level ensemble, the top-K models are instantiated with their fine-tuned weights. The ViT models are truncated at the flatten layer. The CNN models are truncated at their deepest convolutional layer and added with a flatten layer. The output from the flattened layers of the ViT and CNN models are then concatenated and appended with the final dense layer to output class probabilities. The weights of trainable layers are frozen and only the final dense layer is trained to output probabilities of classifying the lateral CXRs into normal or TB categories. The model-level ensemble is trained using an SGD optimizer and an initial learning rate of 1e-5. Callbacks are used to store model checkpoints and reduce the learning rate whenever the validation performance did not improve. The best-performing model with the least validation loss is stored to predict the hold-out test set. **Figure 2** illustrates the construction of model-level ensembles using the fine-tuned CNN and ViT models. The performance of the models during CXR modality-specific pretraining, fine-tuning, and ensemble learning are evaluated using the following metrics: (i) accuracy; (ii) area under the receiver-operating-characteristic curve (AUROC); (iii) area under the precision-recall curve (AUPRC); (iv) precision; (v) recall; (vi) F-score; (vii) Matthews correlation coefficient (MCC), (viii) Diagnostic Odds Ratio (DOR), and (ix) Cohen's Kappa. These metrics are expressed in **Eqs 1–11**.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\left((TP + FP)(TP + FN)(TN + FP)(TN + FN)\right)^{1/2}} \tag{5}$$

$$DOR = \frac{(TP \times TN)}{(FP \times FN)} \tag{6}$$

$$P_o = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{7}$$

$$P_{true} = \frac{(TP + FN)(TP + FP)}{(TP + FP + FN + TN)^2} \tag{8}$$

$$P_{false} = \frac{(FP + TN)(FN + TN)}{(TP + FP + FN + TN)^2} \tag{9}$$

$$P_e = P_{true} + P_{false} \tag{10}$$

$$Cohen's\,Kappa = \frac{(P_o - P_e)}{1 - P_e} \tag{11}$$

Here, TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative values, respectively. The models are trained and evaluated using Tensorflow Keras version 2.6.2 on a Linux system with NVIDIA GeForce GTX 1080 Ti GPU, and CUDA dependencies for GPU acceleration.

2.4 Model Explainability

DL models are often criticized for their "black box" behavior, i.e., lack of explanations toward their predictions. This lack of explainability could be attributed to (i) their architectural depth that may not allow decomposability into explainable components and (ii) the presence of non-linear layers that perform complex data transformations and result in non-deterministic behavior that adversely impacts clinical interpretations. Methods have been proposed (Selvaraju et al., 2017) to explain model predictions by highlighting discriminative parts of the image that causes the model to classify the images to their respective categories. In this study, we used class-selective relevance maps (CRM) (Kim et al., 2019) to discriminate image regions used by the fine-tuned CNN models to categorize the CXRs as showing TB-consistent findings. It has been reported that the CRM-based visualization (Kim et al., 2019) outperformed the conventional gradient-based class activation maps (Selvaraju et al., 2017) in interpreting model predictions.

We computed the attention maps from the fine-tuned ViT models using the attention rollout method discussed in (Zhai et al., 2021). The steps involved in computing the attention map consists of (i) getting the attention weights from each transformer block, (ii) averaging the attention weights across all the heads, (iii) adding an identity matrix to the attention matrix to account for residual connections, (iv) re-normalizing the weights and recursively multiplying the weight matrices to mix the attention across tokens through all the layers, and (v) computing the attention from the output token to the input space. The bounding box coordinates of the heatmaps and attention maps are computed as follows: (i) A difference binary image is generated using the original input lateral CXR image and the heatmap/attention map-overlaid image; (ii) the polygonal coordinates of the connected components in the binary image are measured that gives the coordinates of the vertices and that of the line segments making up the sides of the polygon, and (iii) a binary mask is generated from the polygon and the coordinates are stored for further analysis. The delineated ROIs are compared against the ground truth annotations provided by the radiologist.

For evaluating localization performance, we used several ensemble methods, such as simple averaging, SLSQP-based weighted averaging, and a bitwise-AND of the heatmaps and attention maps of top-K performing models. In simple averaging, the heatmaps and attention maps obtained respectively using the CNN and ViT models are averaged to produce the final heatmap, highlighting discriminative ROIs toward TB detection. In SLSQP-based weighted averaging, the optimal weights obtained using the SLSQP method are used while averaging the heatmaps and attention maps. In a bitwise-AND ensemble, the heatmaps and attention maps are binarized and bitwise-ANDed. The corresponding pixel in the final heatmap is activated only if there is complete agreement among activations in the candidate heatmaps and attention maps. The ROI localization performance of the constituent models and their ensembles is measured in terms of the mean average precision (mAP) metric. The mAP is calculated by taking the mean precision over 11 IoU threshold values within the range [0.1, 0.6] at equal intervals of 0.05 [denoted as mAP@[0.1, 0.6]] (GTUA et al., 2014).

2.5 Statistical Significance Analysis

It has been reported in (Diong et al., 2018) that 90–96% of the studies published in scientific journals do not measure statistical significance in the reported results, casting doubt on algorithm reliability and confidence. In this study, we analyzed statistical significance using the 95% confidence intervals (CIs) for the MCC metric measured as the Clopper–Pearson binomial CI interval. For RoI localization, we measured the 95% CIs measured as the Clopper–Pearson binomial CI interval for the mAP metric achieved by the individual models and their ensembles to report statistical significance. The StatsModels and SciPy Python packages are used in this analysis. We obtained the p-value from the CIs using the methods reported in (Altman and Bland, 2011). Considering the upper and lower limits of the 95% CI as u and l respectively, the standard error (SE) is measured as given in **Eq. 12**.

$$SE = (u - l)(2 \times 1.96) \tag{12}$$

The test statistic z is given by **Eq. 13**

$$z = \frac{Diff}{SE} \tag{13}$$

Here, *Diff* denotes the estimated differences between the models for the measured metric.

The p-value is then calculated as given in **Eq 14**.

$$p = exp(-0.717 \times z - 0.416 \times z^2) \tag{14}$$

3 RESULTS

3.1 CXR Modality-Specific Pretraining and Fine-Tuning

Recall that the CNN and ViT models are instantiated with their ImageNet-pretrained weights and retrained on a combined selection of lateral CXRs from the CheXpert and PadChest datasets. The test performance achieved during CXR modality-specific pretraining is shown in **Table 2**. From **Table 2**, we observed the following: (i) The training time for CNN models is comparatively small than ViT models. The EfficientNet-V2-B0 model took the least while the ViT-L/16 model took the most time for training and convergence. (ii) The VGG-16 model demonstrated superior performance in terms of accuracy, F-score, MCC, DOR, Kappa, AUROC, and AUPRC metrics. The EfficientNet-V2-B0 model demonstrated superior recall and ViT-B/32 demonstrated superior precision compared to other models. However, considering a balanced measure of precision and recall, as provided by the MCC metric, the VGG-16 model demonstrated superior performance compared to other models. (iii) We observed that the 95% CIs obtained for the MCC metric using the VGG-16 model are not significantly different ($p > 0.05$) from other models. Due to this lack of statistical significance, all modality-specific pretrained models are fine-tuned to evaluate performance in the TB classification task. **Table 3** shows the performance achieved by the fine-tuned models that classify the lateral CXRs as showing no abnormalities or other abnormalities that are consistent with TB.

The following are observed from **Table 3**: (i) The CNN models took comparatively lesser time to converge than the ViT models. This observation is analogous to CXR modality-specific pretraining. (ii) The DenseNet-121 model demonstrated

TABLE 2 | Test performance achieved by the CNN and ViT models during lateral CXR modality-specific pretraining. The values in parenthesis denote the 95% CI measured as the Clopper–Pearson binomial interval for the MCC metric. Bold numerical values denote superior performance.

Model	Accuracy	Recall	Precision	F	MCC	DOR	Kappa	AUROC	AUPRC	Training time (seconds)
ViT-B/16	0.7747	0.7988	0.8913	0.8425	0.4596 (0.3647,0.5545)	9	0.4512	0.8276	0.9334	17582.14
ViT-B/32	0.7394	0.7151	**0.9218**	0.8054	0.4621 (0.3671,0.5571)	**11**	0.4293	0.8375	0.9375	10739.29
ViT-L/16	0.7678	0.7846	0.8946	0.8360	0.4555 (0.3606,0.5504)	9	0.4442	0.8276	0.9332	54949.73
ViT-L/32	0.7872	0.8324	0.8792	0.8552	0.4584 (0.3635,0.5533)	9	0.4560	0.8364	0.9373	28797.83
EfficientNet-V2-B0	0.7794	**0.8391**	0.8645	0.8516	0.4231 (0.3290,0.5172)	8	0.4223	0.8152	0.9281	2296.54
VGG-16	**0.8009**	0.8361	0.8931	**0.8637**	**0.4998 (0.4046,0.5950)**	**11**	**0.4960**	**0.8526**	**0.9441**	9316.52
DenseNet-121	0.7886	0.8230	0.8885	0.8545	0.4747 (0.3796,0.5698)	10	0.4701	0.8401	0.9393	7281.22

TABLE 3 | Performance achieved by the fine-tuned models toward the TB classification task. The values in parenthesis denote the 95% CI measured as the Clopper–Pearson binomial interval for the MCC metric. Bold numerical values denote superior performance.

Model	Accuracy	Recall	Precision	F	MCC	DOR	Kappa	AUROC	AUPRC	Training time (seconds)
ViT-B/16	0.7642	0.6792	0.8182	0.7422	0.5361 (0.4411,0.6311)	12	0.5283	0.8548	0.8668	828.30
ViT-B/32	0.8302	0.7547	0.8889	0.8163	0.6680 (0.5783,0.7577)	30	0.6604	0.9227	0.9351	338.46
ViT-L/16	0.8302	**0.8302**	0.8302	0.8302	0.6604 (0.5702,0.7506)	24	0.6604	0.8943	0.9084	1539.06
ViT-L/32	0.7736	0.7170	0.8085	0.7600	0.5507 (0.4560,0.6454)	12	0.5472	0.8786	0.8911	574.24
EfficientNet-V2-B0	0.8019	0.6981	0.8810	0.77900	0.6172 (0.5246,0.7098)	22	0.6038	0.8896	0.9025	114.89
VGG-16	0.8208	0.7358	0.8864	0.8041	0.6510 (0.5602,0.7418)	27	0.6415	0.9110	0.9219	267.40
DenseNet-121	**0.8585**	0.8113	**0.8958**	**0.8515**	**0.7202 (0.6347,0.8057)**	**41**	**0.7170**	**0.9288**	**0.9423**	313.44

superior performance in terms of accuracy, precision, F-score, MCC, DOR, Kappa, AUROC, and AUPRC metrics. The ViT-L/16 model demonstrated superior recall compared to other models. However, considering the MCC metric, the DenseNet-121 model demonstrated superior performance compared to other models. (iii) The 95% CIs for the MCC metric achieved

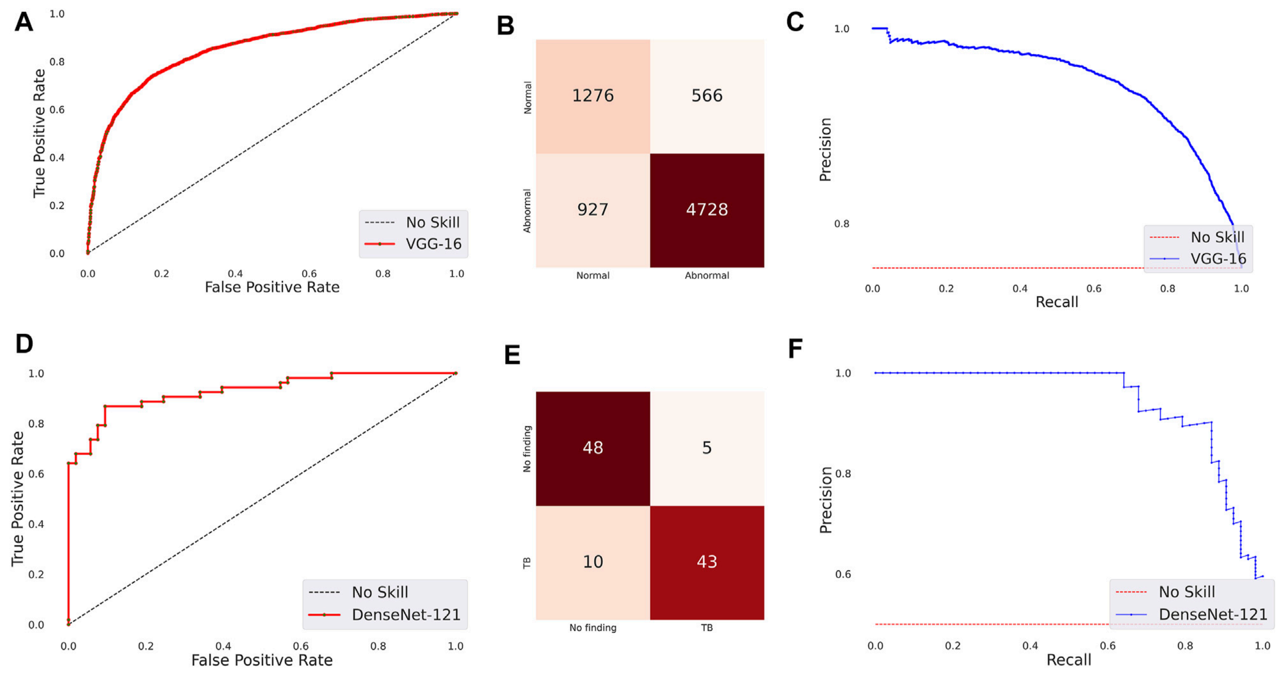


FIGURE 3 | Performance curves achieved by the models used in this study. CXR modality-specific pretraining (VGG-16): **(A)** AUROC; **(B)** AUPRC; **(C)** Confusion matrix. Fine-tuning (DenseNet-121): **(D)** AUROC; **(E)** AUPRC, and **(F)** Confusion matrix.

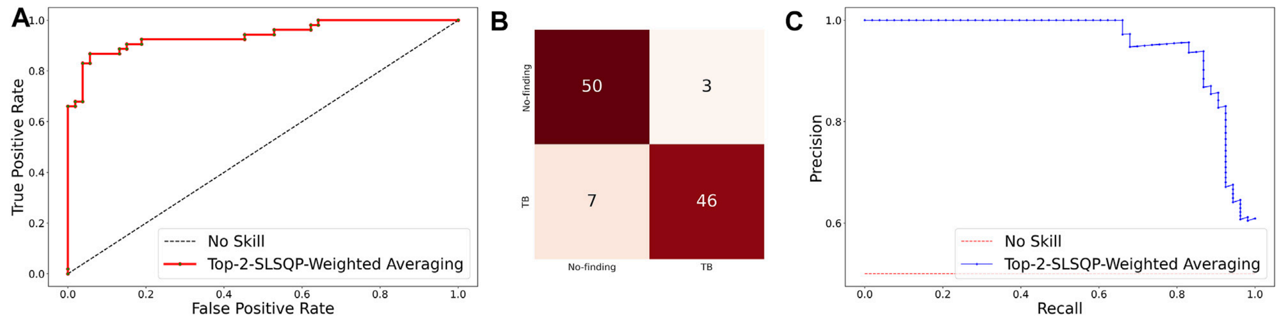


FIGURE 4 | Performance curves achieved using SLSQP-based weighted averaging of the predictions of top-2 fine-tuned models, i.e., DenseNet-121, and ViT-B/32 models. **(A)** AUROC; **(B)** Confusion matrix, and **(C)** AUPRC.

by the DenseNet-121 model demonstrated a tighter error margin, hence higher precision, compared to other models. We observed that the MCC metric achieved by the DenseNet-121 model is significantly superior to ViT-B/16 (p = 0.0001), ViT-L/32 (p = 0.0002), and EfficientNet-V2-B0 (p = 0.0183) models. We also observed that the MCC metric achieved by the VGG-16 model is significantly superior to the ViT-B/16 (p = 0.0133) and ViT-L/32 (p = 0.0304) models. These observations underscore the fact that the CNN models delivered superior classification performance compared to the ViT models. **Figure 3** shows the AUROC, AUPRC, and confusion matrices achieved by the VGG-16 and DenseNet-121 models during the CXR modality-specific pretraining and fine-tuning stages, respectively. A no-skill classifier fails to discriminate between the classes and would predict a random or a constant class in all circumstances.

The ensemble of the top-K models (K = 2, 3, 5, 7) is constructed to evaluate any improvement in classification performance during fine-tuning. **Table 4** shows the performance achieved using various ensemble methods discussed in this study. From **Table 4**, we observe that the performance obtained through SLSQP-based weighted averaging is comparatively higher than other ensembles and their constituent models. This demonstrates that, unlike using equal weights, the use of optimal weights to combine the predictions of constituent models improved classification performance. (ii) The SLSQP-based weighted averaging [optimal weights: (0.65, 0.35)] of the predictions of the top-2 fine-tuned models, viz. DenseNet-121 and ViT-B/32 delivered superior performance in terms of accuracy, Kappa, and significantly superior performance in terms of the MCC metric (0.8136, 95% CI: (0.7394, 0.8878)) compared to its constituent models, viz. DenseNet-121 (p = 0.0137), and ViT-B/32 (p = 0.0002). This ensemble also demonstrated significantly superior performance in terms of MCC metric compared to other models, viz. VGG-16 (p = 0.0001), EfficientNet-V2-B0 (p = 0.0001), ViT-B/16 (p = 0.0001), ViT-L/16 (p = 0.0001),

TABLE 4 | Test performance obtained using prediction-level and model-level ensembles. The values in parenthesis denote 95% CI for the MCC metric measured as the Clopper-Pearson binomial interval. Bold numerical values denote superior performance.

Ensemble	Models	Accuracy	Recall	Precision	F-score	MCC	DOR	Kappa	AUROC	AUPRC	Training time (seconds)
Majority voting	Top-2	0.8774	**0.8868**	0.8704	0.8785	0.7549 (0.6730,0.8368)	51	0.7547	0.8774	0.9069	NA
	Top-3	0.8679	0.8302	0.898	0.8628	0.738 (0.6542,0.8218)	47	0.7358	0.8679	0.9065	NA
	Top-5	0.8585	0.7925	0.913	0.8485	0.7233 (0.6381,0.8085)	47	0.717	0.8585	0.9046	NA
	Top-7	0.8585	0.7925	0.913	0.8485	0.7233 (0.6381,0.8085)	47	0.717	0.8585	0.9046	NA
Simple averaging	Top-2	0.8679	0.8113	0.9149	0.86	0.7406 (0.6571,0.8241)	53	0.7358	0.9388	0.9525	NA
	Top-3	0.8491	0.8113	0.8776	0.8431	0.7001 (0.6128,0.7874)	34	0.6981	0.9377	0.9515	NA
	Top-5	0.8679	0.8113	0.9149	0.86	0.7406 (0.6571,0.8241)	53	0.7358	0.937	0.949	NA
	Top-7	0.8396	0.7925	0.875	0.8317	0.6823 (0.5936,0.7710)	30	0.6792	0.9313	0.9441	NA
SLSQP-weighted averaging	Top-2	**0.9057**	0.8679	0.9388	0.902	**0.8136 (0.7394,0.8878)**	110	**0.8113**	0.9409	0.9542	NA
	Top-3	**0.9057**	**0.8868**	0.9216	**0.9039**	0.8119 (0.7375,0.8863)	96	**0.8113**	0.9352	0.9492	NA
	Top-5	0.8962	0.8491	0.9375	0.8911	0.796 (0.7192,0.8728)	94	0.7925	0.9388	0.952	NA
	Top-7	**0.9057**	0.8679	0.9388	0.902	**0.8136 (0.7394,0.8878)**	110	**0.8113**	0.937	0.9503	NA
Model-level	Top-2	0.8962	0.8113	0.9773	0.8866	0.8041 (0.7285,0.8797)	**223**	0.7925	**0.9491**	**0.9587**	91.4263
	Top-3	0.8679	0.7736	**0.9535**	0.8542	0.7493 (0.6667,0.8319)	87	0.7358	0.9274	0.9433	418.05
	Top-5	0.8679	0.7736	**0.9535**	0.8542	0.7493 (0.6667,0.8319)	87	0.7358	0.9427	0.9525	555.088
	Top-7	0.8585	0.7547	0.9524	0.8421	0.7329 (0.6486,0.8172)	79	0.717	0.9366	0.9493	758.957

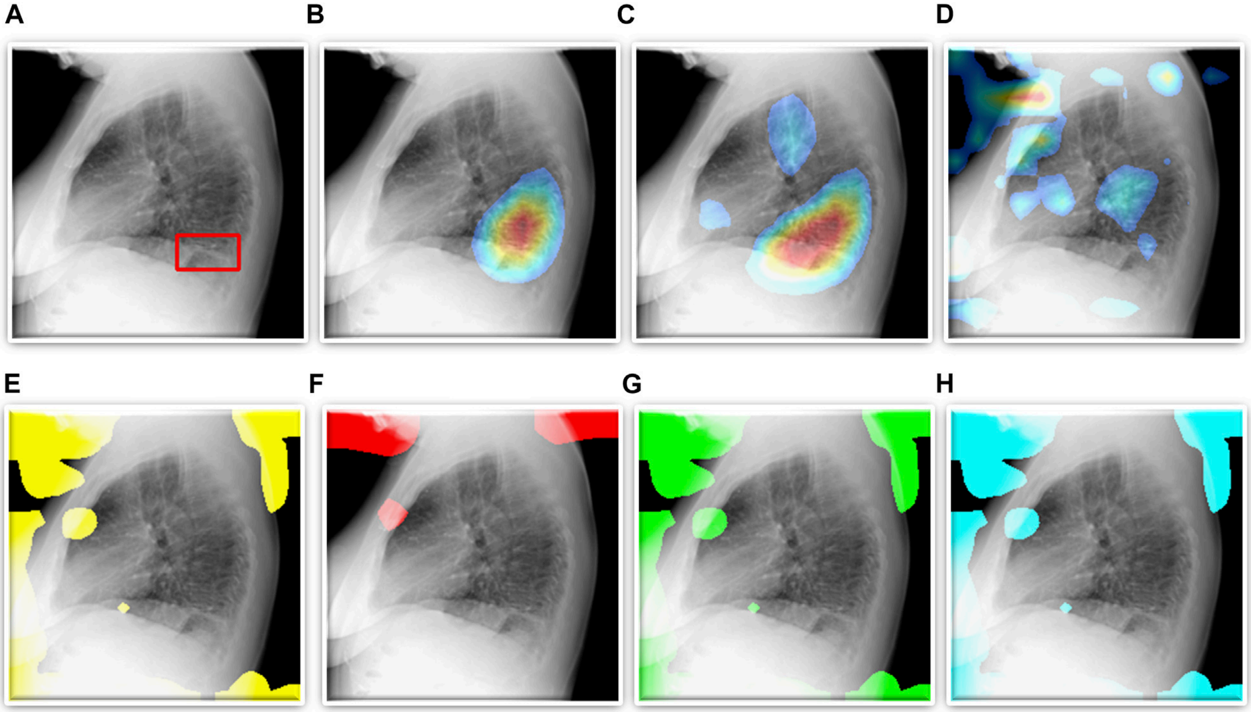


FIGURE 5 | TB-consistent ROI localization achieved using the fine-tuned models. **(A)** An instance of lateral CXR with expert-annotated ROI consistent with TB (shown with a red bounding box); **(B)** VGG-16; **(C)** DenseNet-121; **(D)** EfficientNet-V2-B0; **(E)** ViT-B/16; **(F)** ViT-B/32; **(G)** ViT-L/16, and **(H)** ViT-L/32.

and ViT-L/32 (p = 0.0001) models. The model-level ensemble of the top-2 fine-tuned models, i.e., DenseNet-121 and ViT-B/32 demonstrated superior values for the DOR metric. **Figure 4** shows the AUROC, AUPRC, and confusion matrices achieved by the SLSQP-based weighted averaging of the predictions of the top-2 fine-tuned models.

3.2 Evaluating TB-Consistent ROI Localization Performance

As described in **Section 2.4**, we use CRMs and attention maps to interpret the predictions of the CNN and ViT models, respectively. The delineated ROIs are compared against the ground truth annotations provided by the radiologist. **Figure 5** shows a sample lateral CXR with expert-annotated ROI consistent with TB and the discriminative ROIs highlighted by the fine-tuned CNN and ViT models discussed in this study. **Table 5** shows the TB-consistent ROI localization performance in terms of mAP metric, achieved by the individual models.

TABLE 5 | TB-consistent ROI localization performance achieved by the fine-tuned CNN and ViT models. The values in parenthesis denote the 95% CI measured as the Clopper-Pearson binomial interval for the mAP metric. Bold numerical values denote superior performance.

Model	mAP@[0.1, 0.6]
ViT-B/16	0.0573 (0,0.1205)
ViT-B/32	0.0567 (0,0.1196)
ViT-L/16	0.0573 (0,0.1205)
ViT-L/32	0.0573 (0,0.1205)
EfficientNet-V2-B0	0.0690 (0.0001,0.1379)
VGG-16	**0.1283 (0.0374,0.2192)**
DenseNet-121	0.1052 (0.0218,0.1886)

Further, we constructed ensembles of the heatmaps of the top-2 models from **Table 5**, viz. VGG-16 and DenseNet-121 models using simple averaging, SLSQP-based weighted averaging, and bitwise-AND techniques. **Figure 6** shows the box plots for the range of mAP values achieved by the individual models and other ensembles. **Table 6** shows the TB-consistent ROI localization performance achieved in terms of the mAP metric by the model ensembles.

From **Figure 6**, we observe that the maximum, mean, median, the total range, and the inter-quartile range of the mAP values achieved with the Bitwise-AND ensemble is significantly higher (p < 0.05) than those obtained with the ViT models and considerably higher than the averaging and weighted averaging ensembles. From **Table 6**, we observe that all ensemble methods demonstrated superior values for the mAP metric compared to the individual models (**Table 5**). The bitwise-AND operation resulted in superior values for the mAP metric compared to the constituent models, other models, and ensembles. The mAP metric achieved by the bitwise-AND ensemble is observed to be significantly superior to ViT-B/16, ViT-L/16, ViT-L/32 (p = 0.0199), ViT-B/32 (p = 0.0193), and EfficientNet-V2-B0 (p = 0.0014) models. This performance is followed by the SLSQP-based weighted averaging ensemble that demonstrated significantly

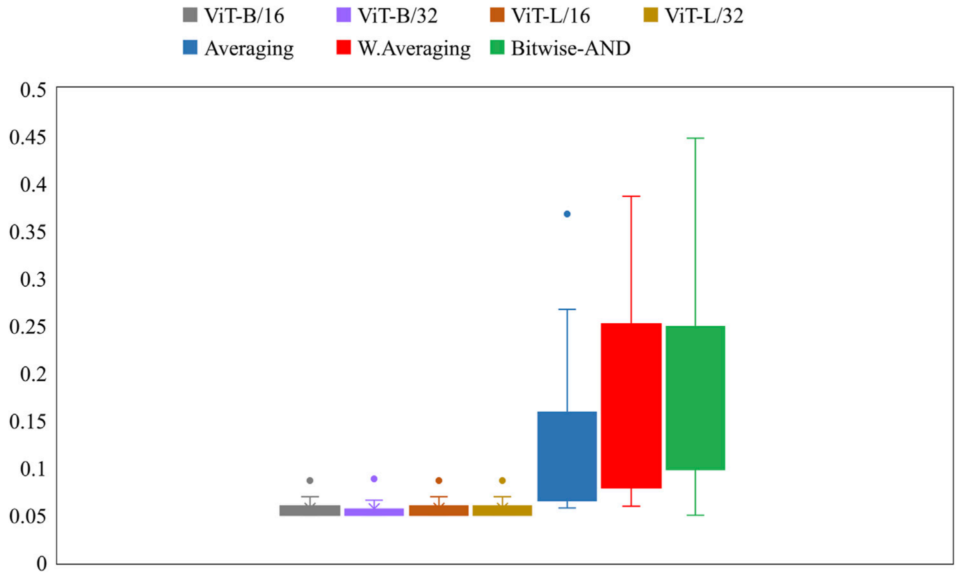


FIGURE 6 | Box plots showing the range of mAP values obtained by the individual models and other ensembles.

TABLE 6 | TB-consistent ROI localization performance achieved by the model ensembles. The values in parenthesis denote the 95% CI measured as the exact Clopper-Pearson binomial interval for the mAP metric. Bold numerical values denote superior performance.

Model	mAP@[0.1, 0.6]
Simple averaging	0.1332 (0.0408,0.2256)
SLSQP-weighted averaging	0.1742 (0.0711,0.2773)
Bitwise-AND	**0.1820 (0.0771,0.2869)**

superior localization performance compared to ViT-B/16, ViT-L/16, ViT-L/32 (p = 0.0264), and EfficientNet-V2-B0 (p = 0.0029) models. **Figure 7** shows a Bitwise-AND ensemble of the heatmaps produced by the top-2 models, viz. VGG-16 and DenseNet-121 models, for instances of test images.

4 DISCUSSION

Following findings from our pilot studies which are consistent with prior observations [34], the ImageNet-pretrained CNNs with their total depth and the ImageNet-pretrained ViT models demonstrated sub-optimal performance toward the task of TB detection. Therefore, we truncated the ImageNet-pretrained CNN models at their optimal intermediate layers, appended them with the classification layers. Further, instead of using ImageNet weights learned from stock photographic images we trained the CNN and ViT models on a large-scale collection of lateral CXR data. These CXR modality-specific pretrained weights serve as a promising initialization to promote modality-specific knowledge transfer and improved adaptation and performance of the models in the relevant task of detecting TB-consistent manifestations.

From our findings and evaluation results, we observe that the ViT models demonstrate sub-optimal classification and ROI localization performance and significantly higher training time, compared to the CNN-based DL models. These findings confirm our suspicion that these may be due to the lack of intrinsic inductive biases. On the other hand, CNN models show superior performance at lower training times even with our limited dataset. Even though CheXpert and PadChest data sets have a cumulative of over 384,316 CXRs only 76,033 lateral CXRs are found in them with only 530 lateral CXRs (0.13% of the total number of lateral CXRs) exhibiting manifestations consistent with TB. This could be a significant factor in the sub-optimal performance exhibited by the ViT models. We improved both classification and ROI localization performance, qualitatively and quantitatively, using CXR modality-specific training, fine-tuning, and constructing model ensembles. This performance improvement with ensemble learning is consistent with the literature (He et al., 2016; Rajaraman et al., 2018a; Rajaraman et al., 2019).

We also show that classification performance is not indicative of reliable disease prediction. For example, even though the average classification performance of ViT models is approximately 80%, their average MAP score is only 5.7% which is evident from the visualization studies, examples of which are shown in **Figures 5E–H**. This underscores the need for visualization of localized disease prediction regions to verify model credibility.

Regarding the use of ensembles, we find in the literature a frequent use of methods such as majority voting, simple averaging, and weighted averaging with equal eights. However, we show that using optimized weighting using

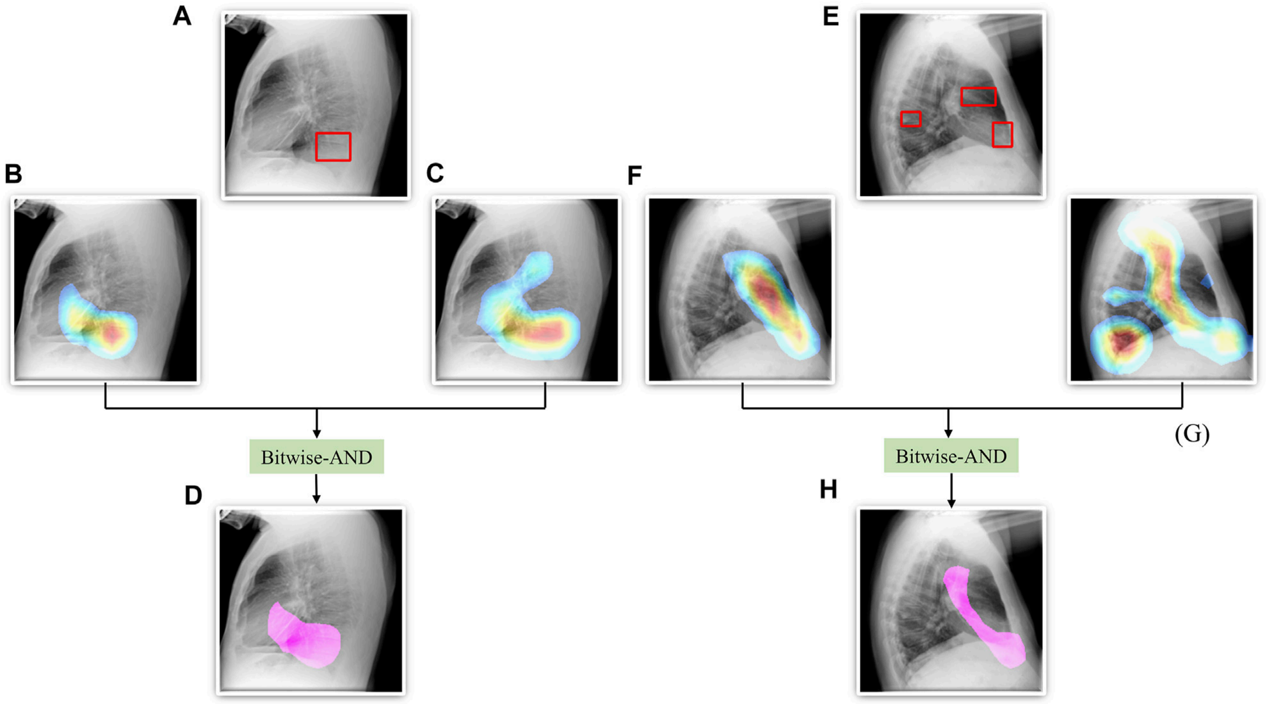


FIGURE 7 | A Bitwise-AND ensemble generated using the heatmaps produced by the top-2 performing models, viz. VGG-16 and DenseNet-121 models. **(A)** and **(E)** Sample test lateral CXRs with expert ground truth annotations (shown in red bounding box); **(B)** and **(F)** Heatmaps produced by the VGG-16 model; **(C)** and **(G)** Heatmaps produced by the DenseNet-121 model, and **(D)** and **(H)** Mask resulting from the Bitwise-AND operation of the heatmaps produced by the VGG-16 and DenseNet-121 models.

specialized techniques, such as SLSQP, result in significantly superior classification performance, e.g., the SLSQP accuracy achieved with the top-2 models is 0.9057 compared to 0.8679 for simple averaging ($p = 0.0001$). Similar behavior is observed for localization performance as well.

Our study has the following limitations: (i) Lateral CXRs help confirm abnormal opacification spatial location, however, have more overlapping structures (e.g., shoulders including scapula and humeral heads), decreasing conspicuity relative to frontal projections. Given that there are more frontal projection CXRs available with TB manifestations, we provide an avenue to explore the combination including lateral images that we believe will improve performance. (ii) There are a very small number of lateral CXRs with TB-consistent findings available for fine-tuning the models which have, very likely, affected the sub-par performance of ViT models as they demand more training data and training time due to their functional characteristics. We expect that the performance of the models would scale with increased data and appropriate empowerment of computational resources. (iii) There is also an imbalance in the number of left or right lateral CXRs in an already small dataset of 530 TB disease-positive images. On the positive side, through augmentation, ensemble learning, and optimized weighting of model predictions, we were able to achieve a lateral-view agnostic performance that was significantly high. However, it is important to consider that the anatomical view presented in a left lateral image is different from the one presented in the other. For clinical diagnostic or screening applications, it would be necessary to train the classifier on these differences so that a reliable and robust interpretation of the prediction can be obtained. Further, research is ongoing in building combination model architectures like ConViT (d'Ascoli et al., 2021) that combines characteristics of the CNN and ViT architectures toward improving performance. Such models should be studied in future studies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SR: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review and editing; GZ: Conceptualization, Methodology, Writing—review and editing; LF: Data curation (lateral CXR annotations), Writing—review and editing; SA: Conceptualization, Formal analysis, Funding acquisition, Investigation, Project administration, Resources, Supervision, Validation, Writing—review and editing.

REFERENCES

Altman, D. G., and Bland, J. M. (2011). How to Obtain the P Value from a Confidence Interval. *BMJ* 343, d2304. doi:10.1136/bmj.d2304

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions. *J. Big Data* 8, 53. doi:10.1186/s40537-021-00444-8

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. (2020). PadChest: A Large Chest X-ray Image Dataset with Multi-Label Annotated Reports. *Med. Image Anal.* 66, 101797. doi:10.1016/j.media.2020.101797

d'Ascoli, S., Touvron, H., Leavitt, M., Morcos, A., Biroli, G., and Sagun, L. (2021). *ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases*. Available from: http://arxiv.org/abs/2103.10697.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. *Mult. Classif Syst.* 1857, 1–15. doi:10.1007/3-540-45014-9_1

Ding, M., Antani, S., Jaeger, S., Xue, Z., Candemir, S., Kohli, M., et al. (2017). "Local-global Classifier Fusion for Screening Chest Radiographs," in *Proc. Of SPIE Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*. Editors J. Z. Tessa and S. Cook. doi:10.1117/12.2252459

Diong, J., Butler, A. A., Gandevia, S. C., and Héroux, M. E. (2018). Poor Statistical Reporting, Inadequate Data Presentation and Spin Persist Despite Editorial Advice. *PLoS One* 13, e0202121. doi:10.1371/journal.pone.0202121

Duong, L. T., Le, N. H., Tran, T. B., Ngo, V. M., and Nguyen, P. T. (2021). Detection of Tuberculosis from Chest X-ray Images: Boosting the Performance with Vision Transformer and Transfer Learning. *Expert Syst. Appl.* 184, 115519. doi:10.1016/j.eswa.2021.115519

Dutta, A., and Zisserman, A., 2019. The VIA Annotation Software for Images, Audio and Video. in MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia October 2019 2276–2279. doi:10.1145/3343031.3350535

Gaber, K. A., McGavin, C. R., and Wells, I. P. (2005). Lateral Chest X-ray for Physicians. *J. R. Soc. Med.* 98, 310–312. doi:10.1258/jrsm.98.7.31010.1177/014107680509800705

Gtua, Colleges., Academy, O., Academy, O., Academy, O., Science, A. C., Technology, I., et al. (2014). "Microsoft COCO," in *European Conference on Computer Vision*, 740–755.

Gupta, M., and Gupta, B. (2018).An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP), Proceeding of the 2018 11th Int Conf Contemp Comput IC3 2018, Aug. 2018, Noida, India. IEEE, 2–4. doi:10.1109/IC3.2018.8530572

He, K., Zhang, X., Ren, S., and Sun, J. (2016).Deep Residual Learning for Image Recognition, Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016, Las Vegas, NV, USA. IEEE, 770–778. doi:10.1109/CVPR.2016.90

Herrera Diaz, M., Haworth-Brockman, M., and Keynan, Y. (2020). Review of Evidence for Using Chest X-Rays for Active Tuberculosis Screening in Long-Term Care in Canada. *Front. Public Health* 8, 8. doi:10.3389/fpubh.2020.00016

Hogeweg, L., Mol, C., De Jong, P. A., Dawson, R., Ayles, H., and Van Ginneken, B. (2010). "Fusion of Local and Global Detection Systems to Detect Tuberculosis in Chest Radiographs," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*, 650–657. doi:10.1007/978-3-642-15711-0_81

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017).Densely Connected Convolutional Networks, Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, July 2017, Honolulu, HI, USA. IEEE. doi:10.1109/CVPR.2017.243

FUNDING

This study is supported by the Intramural Research Program (IRP) of the National Library of Medicine (NLM) and the National Institutes of Health (NIH).

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., et al. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *Proc. AAAI Conf. Artif. Intelligence* 33, 590–597. doi:10.1609/aaai.v33i01.3301590

Islam, M. T., Aowal, M. A., Minhaz, A. T., and Ashraf, K. (2017). *Abnormality Detection and Localization in Chest X-Rays Using Deep Convolutional Neural Networks*. New York, NY: arXiv. Available from: http://arxiv.org/abs/1705.09850.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y. X., Lu, P. X., and Thoma, G. (2014). Two Public Chest X-ray Datasets for Computer-Aided Screening of Pulmonary Diseases. *Quant Imaging Med. Surg.* 4, 475–477. doi:10.3978/j.issn.2223-4292.2014.11.20

Kim, I., Rajaraman, S., and Antani, S. (2019). Visual Interpretation of Convolutional Neural Network Predictions in Classifying Medical Image Modalities. *Diagnostics* 9, 38. doi:10.3390/diagnostics9020038

Lakhani, P., and Sundaram, B. (2017). Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 284, 574–582. doi:10.1148/radiol.2017162326

Liu, C., and Yin, Q. (2021). Automatic Diagnosis of COVID-19 Using a Tailored Transformer-like Network. *J. Phys. Conf. Ser.* 2010, 012175. doi:10.1088/1742-6596/2010/1/012175

Park, S., Kim, G., Oh, Y., Seo, J. B., Lee, S. M., Kim, J. H., et al. (2022). Multi-task Vision Transformer Using Low-Level Chest X-ray Feature Corpus for COVID-19 Diagnosis and Severity Quantification. *Med. Image Anal.* 75, 102299. doi:10.1016/j.media.2021.102299

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., and Pfeiffer, D. (2019). Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci. Rep.* 9, 6268. doi:10.1038/s41598-019-42557-4

Rajaraman, S., and Antani, S. K. (2020). Modality-Specific Deep Learning Model Ensembles toward Improving TB Detection in Chest Radiographs. *IEEE Access* 8, 27318–27326. doi:10.1109/ACCESS.2020.2971257

Rajaraman, S., Candemir, S., Xue, Z., Alderson, P. O., Kohli, M., Abuya, J., et al. (2018).A Novel Stacked Generalization of Models for Improved TB Detection in Chest Radiographs, Conf Proc . Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf, July 2018, Honolulu, HI, USA. IEEE, 718–721. doi:10.1109/EMBC.2018.8512337

Rajaraman, S, Rajaraman, S., Candemir, S., Kim, I., Thoma, G., Antani, S., et al. (2018). Visualization and Interpretation of Convolutional Neural Network Predictions in Detecting Pneumonia in Pediatric Chest Radiographs. *Appl. Sci.* 8, 1715. doi:10.3390/app8101715

Rajaraman, S., Sornapudi, S., Kohli, M., and Antani, S. (2019).Assessment of an Ensemble of Machine Learning Models toward Abnormality Detection in Chest Radiographs, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, July 2019, Berlin, Germany. IEEE. doi:10.1109/EMBC.2019.8856715

Rajaraman, S., Siegelman, J., Alderson, P. O., Folio, L. S., Folio, L. R., and Antani, S. K. (2020). Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays. *IEEE Access* 8, 115041–115050. doi:10.1109/ACCESS.2020.3003810

Rajaraman, S., Kim, I., and Antani, S. K. (2020). Detection and Visualization of Abnormality in Chest Radiographs Using Modality-specific Convolutional Neural Network Ensembles. *PeerJ* 8, e8693. doi:10.7717/peerj.8693

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., et al. (2017). *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning*.

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks* 61, 85–117. doi:10.1016/j.neunet.2014.09.003

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017).Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Proceedings of the IEEE International Conference on Computer Vision, 2017-Oct, Venice, Italy. IEEE, 618–626. doi:10.1109/ICCV.2017.74

Shome, D., Kar, T., Mohanty, S., Tiwari, P., Muhammad, K., Altameem, A., et al. (2021). Covid-transformer: Interpretable Covid-19 Detection Using Vision Transformer for Healthcare. *Int. J. Environ. Res. Public Health* 18, 11086. doi:10.3390/ijerph182111086

Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. in Proceeding of the 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015,april 2015

Sivaramakrishnan, R., Antani, S., Candemir, S., Xue, Z., Thoma, G., Alderson, P., et al. (2018). "Comparing Deep Learning Models for Population Screening Using Chest Radiography," in *SPIE Medical Imaging* (Houston, Texas, United States): SPIE). doi:10.1117/12.2293140

Swingler, G. H., Du Toit, G., Andronikou, S., Van Der Merwe, L., and Zar, H. J. (2005). Diagnostic Accuracy of Chest Radiography in Detecting Mediastinal Lymphadenopathy in Suspected Pulmonary Tuberculosis. *Arch. Dis. Child.* 90, 1153–1156. doi:10.1136/adc.2004.062315

Tan, M., and Le, Q. V. (2021). *EfficientNetV2: Smaller Models and Faster Training*. Available at: http://arxiv.org/abs/2104.00298.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. (2017).ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, Proceeding of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017, Honolulu, HI, USA. IEEE, 1–19. doi:10.1109/CVPR.2017.369

World Health Organization (2016). *Chest Radiography in Tuberculosis*. Switzerland: World Heal Organ, 1–44. Available at: https://apps.who.int/iris/handle/10665/252424.

Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). "The Ninth Annual Conference on Learning Representations (ICLR 2021)," in A Virtual Conference Due to COVID-19 Pandemic, Vienna, May 3–7.

Conflict of Interest: LF has two issued patents (no royalties since NIH and military-owned) related to chest imaging: (i) "Radiographic marker that displays upright angle on portable x-rays," US Patent 9,541,822 B2, and (ii) "Multigrayscale universal CT Window," US Patent 8,406,493 B2.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 March 2022
doi: 10.3389/fgene.2022.848116

A Joint Model of Random Forest and Artificial Neural Network for the Diagnosis of Endometriosis

Jiajie She[1,2], *Danna Su*[1], *Ruiying Diao*[1]* *and Liping Wang*[1]*

[1]*Reproductive Medicine Centre, Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen, China,* [2]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China*

Endometriosis (EM), an estrogen-dependent inflammatory disease with unknown etiology, affects thousands of childbearing-age couples, and its early diagnosis is still very difficult. With the rapid development of sequencing technology in recent years, the accumulation of many sequencing data makes it possible to screen important diagnostic biomarkers from some EM-related genes. In this study, we utilized public datasets in the Gene Expression Omnibus (GEO) and Array-Express database and identified seven important differentially expressed genes (DEGs) (*COMT*, *NAA16*, *CCDC22*, *EIF3E*, *AHI1*, *DMXL2*, and *CISD3*) through the random forest classifier. Among these DEGs, *AHI1*, *DMXL2*, and *CISD3* have never been reported to be associated with the pathogenesis of EMs. Our study indicated that these three genes might participate in the pathogenesis of EMs through oxidative stress, epithelial–mesenchymal transition (EMT) with the activation of the Notch signaling pathway, and mitochondrial homeostasis, respectively. Then, we put these seven DEGs into an artificial neural network to construct a novel diagnostic model for EMs and verified its diagnostic efficacy in two public datasets. Furthermore, these seven DEGs were included in 15 hub genes identified from the constructed protein–protein interaction (PPI) network, which confirmed the reliability of the diagnostic model. We hope the diagnostic model can provide novel sights into the understanding of the pathogenesis of EMs and contribute to the clinical diagnosis and treatment of EMs.

Keywords: endometriosis, random forest, artificial neural network, diagnostic model, diagnostic efficacy

OPEN ACCESS

Edited by:
Tapas Si,
Bankura Unnayani Institute of Engineering, Bankura, India

Reviewed by:
Soumadip Ghosh,
Institute of Engineering and Management (IEM), India
Soumita Seth,
Aliah University, India

****Correspondence:***
Liping Wang
wlp18665070696@163.com
Ruiying Diao
15889753127@163.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *04 January 2022*
Accepted: *28 January 2022*
Published: *08 March 2022*

Citation:
She J, Su D, Diao R and Wang L (2022) A Joint Model of Random Forest and Artificial Neural Network for the Diagnosis of Endometriosis. Front. Genet. 13:848116. doi: 10.3389/fgene.2022.848116

INTRODUCTION

Endometriosis (EM) is an estrogen-dependent inflammatory disorder, which afflicts about 10%–15% of women of childbearing age (Parasar et al., 2017). It is defined as the presence of endometrial-like tissue outside of the uterine cavity, which can lead to chronic pelvic pain, and infertility (Drabble et al., 2021). However, the true prevalence of EMs is uncertain as visual laparoscopy is the gold standard for the diagnosis of EMs (Taylor et al., 2018). At the moment, Sampson's theory of menstrual blood reflux observed in most patients is commonly accepted in the pathophysiology of EMs, while only a small portion will develop into this disease (Burney and Giudice, 2012). However, it could only explain a portion of EMs. Therefore, it's necessary to further investigate a comprehensive understanding of the pathogenesis of EMs and find effective molecular biomarkers to improve the early diagnosis and treatment of EMs.

DNA microarray technology is a high-throughput detection method that can be used to provide gene expression profiles and thus can help to screen disease-related genes and biomarkers (Yoo et al., 2009). With the rapid development of DNA microarray technology, a large amount of high-throughput data has accumulated available on public platforms. However, how to make effective use of these data to screen critical disease-related genes for the diagnosis of EMs is a great challenge. At present, random forest and neural network are widely applied for disease prediction (Yigit and Isik, 2018; Khan et al., 2019; Shaia et al., 2019; Kugunavar and Prabhakar, 2021). Among them, random forest algorithm can perform random sampling to screen the target variables (Schonlau and Zou, 2020) and has high predicted accuracy (Byeon, 2019; Chen et al., 2020). Furthermore, the artificial neural network can be used to evaluate the accuracy of predicted model with divided training and validation datasets (Curchoe et al., 2020). Currently, there are some useful visualization and analysis tools for neural networks, such as NeuralNetTools (Beck, 2018), spiking neuronal networks (Galindo et al., 2020), and Net2Vis (Bauerle et al., 2021).

Therefore, the combination of random forest and artificial neural network would have better classification performance and more meaningful selected features (Kong and Yu, 2018; Tian et al., 2020). In this study, we firstly identified some differentially expressed genes (DEGs) between EMs and normal samples from public datasets in the Gene Expression Omnibus (GEO) database. Through the random forest classifier, we screened these DEGs and obtained seven important DEGs (*COMT*, *NAA16*, *CCDC22*, *EIF3E*, *AHI1*, *DMXL2*, and *CISD3*). Then, we put these seven DEGs into an artificial neural network to construct a novel diagnostic model and verified its diagnostic efficacy in two public datasets (See the detailed process in **Figure 1**). We hope this diagnostic model can provide novel sights into the pathogenesis of EMs and improve the early diagnosis and treatment of EMs.

MATERIALS AND METHODS

Data Download and Processing

The GSE51981, GSE6364, and GSE7307 datasets were downloaded by the R package “GEOquery” (2.60.0) (Davis and Meltzer, 2007) to obtain the expression profile data. Then, the E-MTAB-694 dataset was downloaded through the Array-Express database. The related annotation information including the platforms, the probes, and ID conversion was obtained from the GEO database. When multiple probes corresponded to one gene symbol, the average expression level of multiple probes was used as the expression level of the corresponding gene. ID conversion was conducted with the R package “org.Hs.eg.db” (v3.13.0). Furthermore, the “removeBatchEffect” function in the R package “LIMMA” (v3.48.3) (Ritchie et al., 2015) was used to adjust batch effects, which were evaluated by principal component analysis (PCA).

Differential Expression and Functional Enrichment Analysis

Differential expression analysis was conducted on 77 EM disease and 71 normal samples of the GSE51981 dataset through the Bayesian analysis of the R package “LIMMA”. The log2FoldChange > 1.5 and p-value < 0.05 were set as the threshold of DEGs. The R package “pheatmap” (v1.0.12) was used to perform clustering analysis of DEGs for the heatmap. To explore the biological significance of these DEGs in the pathogenesis of EMs, GO and KEGG pathway enrichment analyses were performed through the R package “clusterProfiler” (v4.1.3) (Wu et al., 2021) to identify significantly enriched GO terms and significantly enriched KEGG pathways with the threshold of p-value < 0.05.

The Construction of Hub Gene Network

The STRING (v11.5) (https://string-db.org/cgi/input.pl) (Szklarczyk et al., 2021) has been widely applied to construct a protein–protein interaction (PPI) network. Based on those DEGs, the “Multiple proteins” option was selected. In the PPI network, the minimum required interaction score was set as “high confidence (0.700)”. Then, the cytoHubba (Chin et al., 2014) was employed to identify hub genes. The eccentricity algorithm was selected and 15 top-ranked genes were chosen as hub genes. Finally, the hub gene network was visualized with Cytoscape (v3. 9.0) (Demchak et al., 2014).

Screening Differentially Expressed Genes With the Random Forest Model

The R package “randomForest” (v4.6.14) (Liaw et al., 2014) was used to construct a random forest model to screen DEGs. The number of random seeds and decision trees was set as 1–5,000 and 3,000 in the random forest classifier originally, respectively. Finally, the number of random seeds and decision trees was set as 4,543 and 219, respectively, which represented higher accuracy of the constructed model and stable model error. The Gini coefficient method was used to obtain the dimensional importance value of all variables from the constructed random forest model. Those DEGs with an importance value greater than 4 were screened as important genes of EMs for subsequent model construction and verification. The R package “pheatmap” was used to perform clustering analysis of the screened important genes for the heatmap in this dataset.

The Construction and Verification of the Artificial Neural Network Model

The GSE6364 dataset downloaded through the R package “GEOquery” was selected as the training set for the construction of the artificial neural network model. After the data normalization, the R package “neuralnet” (v1.44.2) (Fritsch and Guenther, 2016) was used to construct an artificial neural network model of those important variables. The number of hidden neuron layers should be two-thirds of the number of the input layer plus two-thirds of the number of the output layer. Therefore, six hidden layers were set as the model parameter to construct a classification model of EMs through the predicted gene weight information. The R packages “pROC” (v1.18.0) (Robin et al., 2011) and

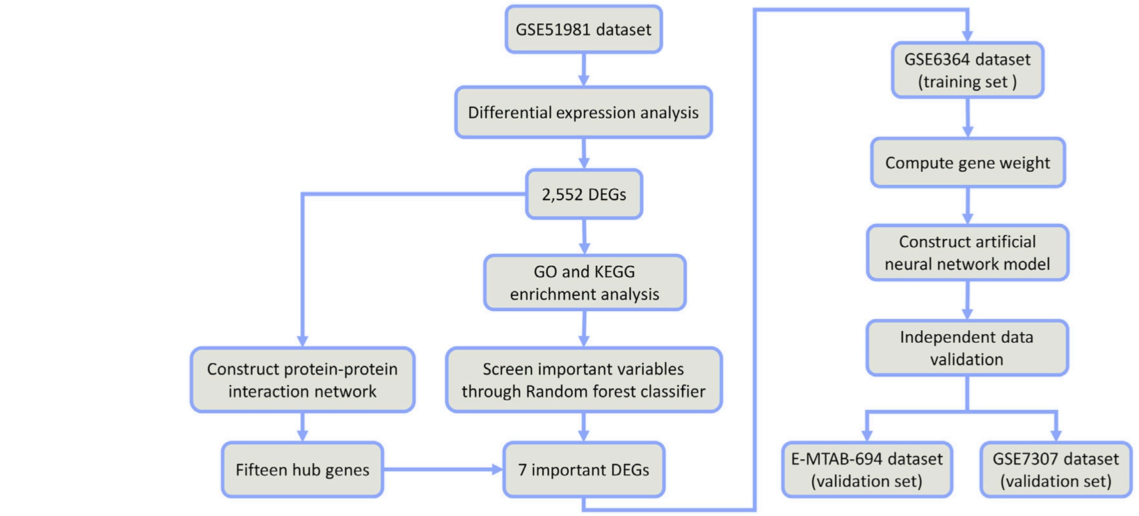

GSE51981 dataset
Differential expression analysis
2,552 DEGs
GO and KEGG enrichment analysis
Construct protein-protein interaction network
Screen important variables through Random forest classifier
Fifteen hub genes
7 important DEGs
GSE6364 dataset (training set)
Compute gene weight
Construct artificial neural network model
Independent data validation
E-MTAB-694 dataset (validation set)
GSE7307 dataset (validation set)

FIGURE 1 | Flow chart.

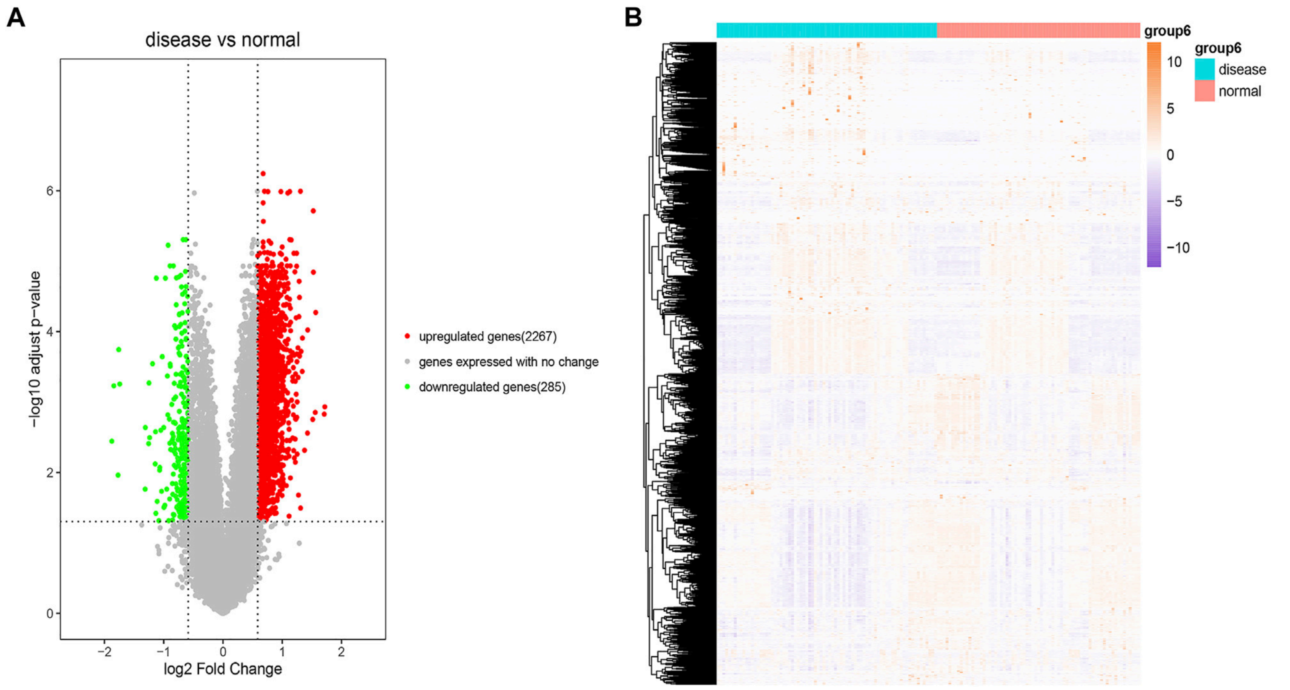


FIGURE 2 | Differential expression analysis. **(A)** Volcano plot of the result of differential expression analysis. The *x*-axis is log2 (fold change) and the *y*-axis is –log10 (adjusted *p*-value). The red dots represent significant upregulated expressed genes. The green dots represent significant downregulated expressed genes. The gray dots represent genes expressed with no change. **(B)** Heatmap of these DEGs. The colors in the graph from red to pink indicate the change from high to low expression levels. On the upper part of the heatmap, the blue band indicates the disease samples and the red band indicates the normal samples.

“ggplot2” (v3.3.5) (Gómez-Rubio, 2017) were used to calculate the verification results of AUC classification performance and draw the ROC curve. Another two datasets E-MTAB-694 and GSE7307 were used to verify the accuracy of the constructed neural network model for the diagnosis of EMs. The R package “pROC” was used to draw ROC curves for each dataset, and the AUC value was calculated to verify the classification efficiency. Meanwhile, the sensitivity and specificity in distinguishing the disease samples from normal samples were calculated.

RESULTS

Data Processing and Differential Expression Analysis

The R package "GEOquery" was used to download the GEO dataset GSE51981 (77 EM disease samples and 71 normal samples) and obtain detailed information. We used the "removeBatchEffect" function in the R package "LIMMA" to adjust batch effects and then conducted principal component analysis (PCA) analysis to evaluate the performance of batch effect adjustment. PCA results (**Supplementary Figure S1**) indicated that the disease samples were mixed with the normal samples, which suggested the challenge of diagnosing. We also used the R package "LIMMA" to perform differential expression analysis for the dataset GSE51981 through the Bayesian test. We finally identified 2,267 significantly upregulated and 285 significantly downregulated expressed genes between the disease samples and the normal samples with the threshold of fold change values of >1.5 and $p < 0.05$. The detailed information of all DEGs is listed in **Supplementary Table S1**. The results of these DEGs and the heatmap of these DEGs are visualized in **Figures 2A** and **2B**, respectively.

Functional Enrichment Analysis for DEGs and the Construction of PPI Network

To explore the biological significance of these DEGs in the pathogenesis of EMs, we performed GO and KEGG pathway enrichment analyses through the R package 'clusterProfiler'. GO terms were classified into three categories: biological process (BP), cellular component (CC), and molecular function (MF). The top five GO terms of genes with significantly upregulated and downregulated expression levels were visualized in **Figures 3A,B**. The GO enrichment analysis results indicated that these significantly upregulated expressed genes were mainly involved in the transmembrane transporter activity, ATPase activity, metallopeptidase activity, aldehyde dehydrogenase NADP$^+$ activity, and lipid transporter activity (**Supplementary Table S2**), while these significantly downregulated expressed genes were mainly involved in the flavin adenine dinucleotide binding, acyl-CoA dehydrogenase activity, phosphatidylcholine transporter activity, extracellular matrix structural constituent, and ATPase-coupled intramembrane lipid transporter activity (**Supplementary Table S3**). For KEGG pathway enrichment analysis, the results indicated that these upregulated expressed genes were significantly associated with the cAMP signaling pathway, adrenergic signaling in cardiomyocytes, aldosterone synthesis and secretion, ABC transporters, and salivary secretion (**Supplementary Table S4**), while these downregulated expressed genes were significantly associated with fatty acid degradation and metabolism; valine, leucine, and isoleucine degradation; lysosome; the PPAR signaling pathway; and the Hippo signaling pathway (**Supplementary Table S5**). Furthermore, we constructed a PPI network through the STRING database. The hub genes selected from the PPI network are shown in **Supplementary Figure S2**. According to the eccentricity scores, we identified 15 hub genes from the network, which had highest confidence scores.

Constructing the Random Forest Model to Screen Differentially Expressed Genes

To screen DEGs, we put these DEGs into the random forest classifier and set the number of random seeds to 4,543. By referring to the relationship between the model error and the number of decision trees (**Figure 4A**), we selected 219 trees as the parameter of the random forest model, which represented a stable error in the model. In the modeling process, we used the Gini coefficient method to measure the importance of all variables according to decreased mean square error and model accuracy (**Figure 4B**). Finally, we selected seven DEGs (*AHI1*, *DMXL2*, *NAA16*, *CCDC22*, *CISD3*, *COMT*, and *EIF3E*) with a mean decrease of Gini index greater than 4 as important variables for subsequent analysis. Interestingly, all these DEGs were included in the 15 hub genes identified from the constructed PPI network. Among these variables, *AHI1* was the most important, with the mean decrease of the Gini index being much higher than other variables (**Supplementary Table S6**). A small number of variables meant a small out-of-band error, which represented a high accuracy of the constructed random forest model. Based on these seven variables, we performed the *k*-means clustering of the dataset. The results suggested that these seven genes could be used to distinguish the disease sample from the normal samples (**Figure 4C**). Furthermore, *AHI1*, *DMXL2*, and *NAA16* genes were clustered as a group with low expression in the normal sample and high expression in the disease sample. On the contrary, *CCDC22*, *CISD3*, *COMT*, and *EIF3E* were clustered as another group with high expression in the normal sample and low expression in the disease sample.

The Construction of the Artificial Neural Network Model and the Evaluation of the ROC Curve

Based on the R package 'neuralnet', we use the GSE6364 dataset (21 disease samples and 21 normal samples) as the training set to construct the artificial neural network model. Firstly, we performed the preprocessing and normalization of this dataset. According to the output results of the neural network model (**Figure 5A**), it is illuminated that the entire training was performed in 11,684 steps. Among the output results, the predicted weights of each hidden neuron layer were −3.97906, 1.04457, 2.76611, −2.00181, −11.84206, and −0.90829 (**Supplementary Table S7**). Next, we drew the ROC curve to evaluate the predicted performance; the AUC values of *AHI1*, *COMT*, *DMXL2*, *CISD3*, *NAA16*, *EIF3E*, and *CCDC22* were 0.7150, 0.7809, 0.6927, 0.7266, 0.7217, 0.7093, and 0.7050, respectively (**Figure 5B**). The larger the AUC value of each DEG is, the higher the credibility of the constructed diagnostic model will be. We also used another two datasets E-MTAB-694 (18 disease samples and 17 normal samples) and GSE7307 (18 disease samples and 23 normal samples) to verify the accuracy of the constructed neural network model. In the E-MTAB-694

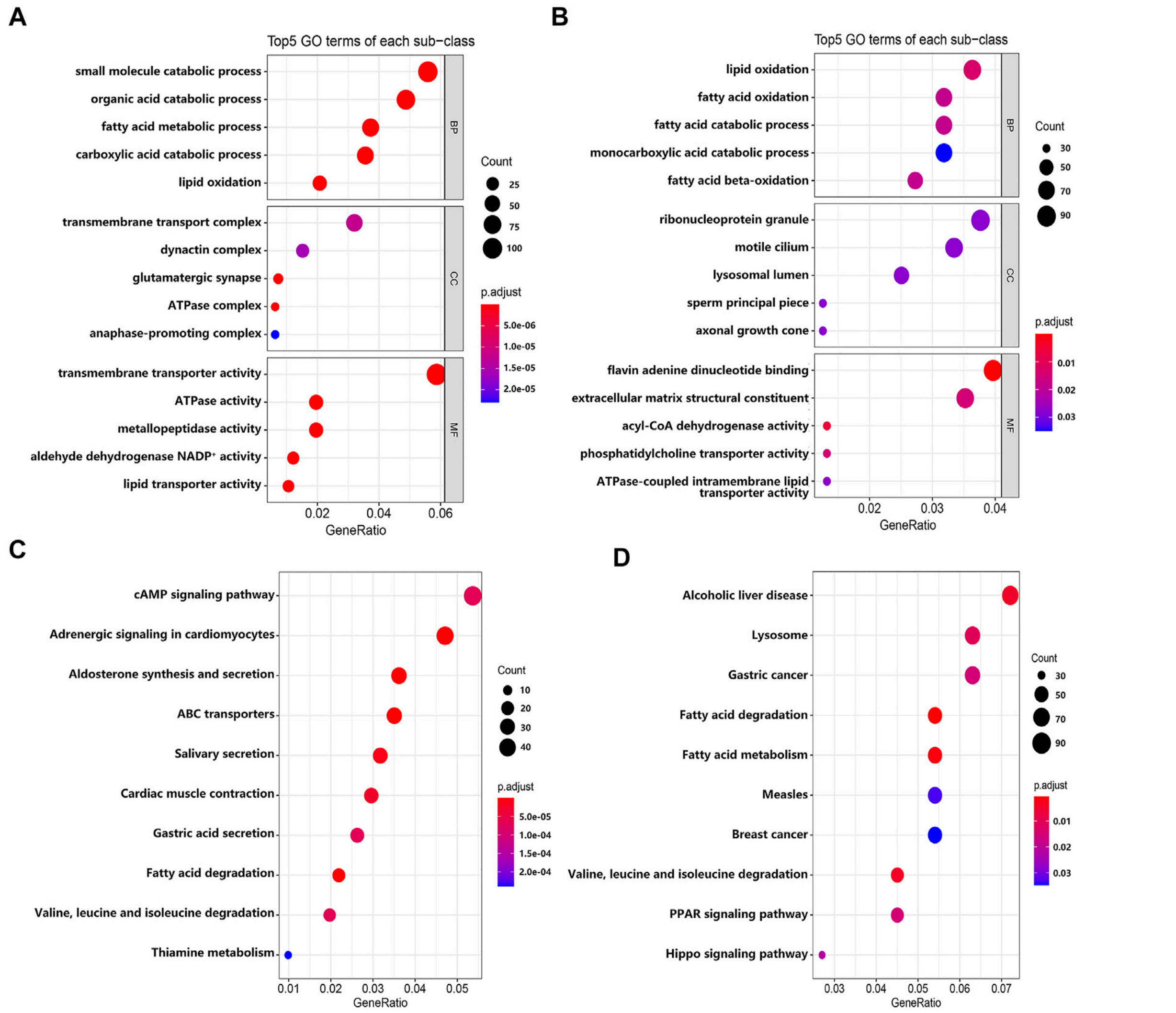


FIGURE 3 | The results of GO and KEGG enrichment analyses. **(A)** The top five GO terms of genes with significantly upregulated expressed level. **(B)** The top five GO terms of genes with significantly downregulated expressed level. **(C)** The top 10 KEGG pathways of genes with significantly upregulated expressed level. **(D)** The top 10 KEGG pathways of genes with significantly downregulated expressed level.

dataset (**Figure 5C**), the AUC values of the seven DEGs were 0.8226, 0.6623, 0.6836, 0.6625, 0.8367, 0.8471, and 0.8617. In the verification results of the GSE7307 dataset (**Figure 5D**), the AUC values of the seven DEGs were 0.7464, 0.6484, 0.7020, 0.6300, 0.9075, 0.8295, and 0.8327. In general, we constructed a novel diagnostic model of EMs and verified its diagnostic efficacy through the constructed artificial neural network in two public datasets.

DISCUSSION

The combination of random forest and artificial neural network can be used to construct a reliable predictive model for the diagnosis of some diseases, such as polycystic ovary syndrome (PCOS) (Xie et al., 2020) and ulcerative colitis (Li et al., 2020). In this study, we identified 2,552 DEGs associated with EMs in the GSE51981 dataset. Based on the random forest classifier, seven important candidate DEGs (*COMT*, *NAA16*, *CCDC22*, *EIF3E*, *AHI1*, *DMXL2*, and *CISD3*) were screened. Then, we used the GSE6364 dataset as the training set to construct the artificial neural network model and evaluated the classification efficacy of the model in E-MTAB-694 and GSE7307 datasets. The AUC values of the ROC curve were about 0.7, which had great efficiency and verified the diagnostic efficacy of the model. Furthermore, we constructed a 15-hub-gene-based PPI network and confirmed the reliability of the prediction model. Compared with the Nnet package, we found that the neuralnet package had higher accuracy of the predicted model (86.5% vs 81.1%). In total, the constructed diagnostic model could provide new insight into our understanding of the pathogenesis of EMs and identify

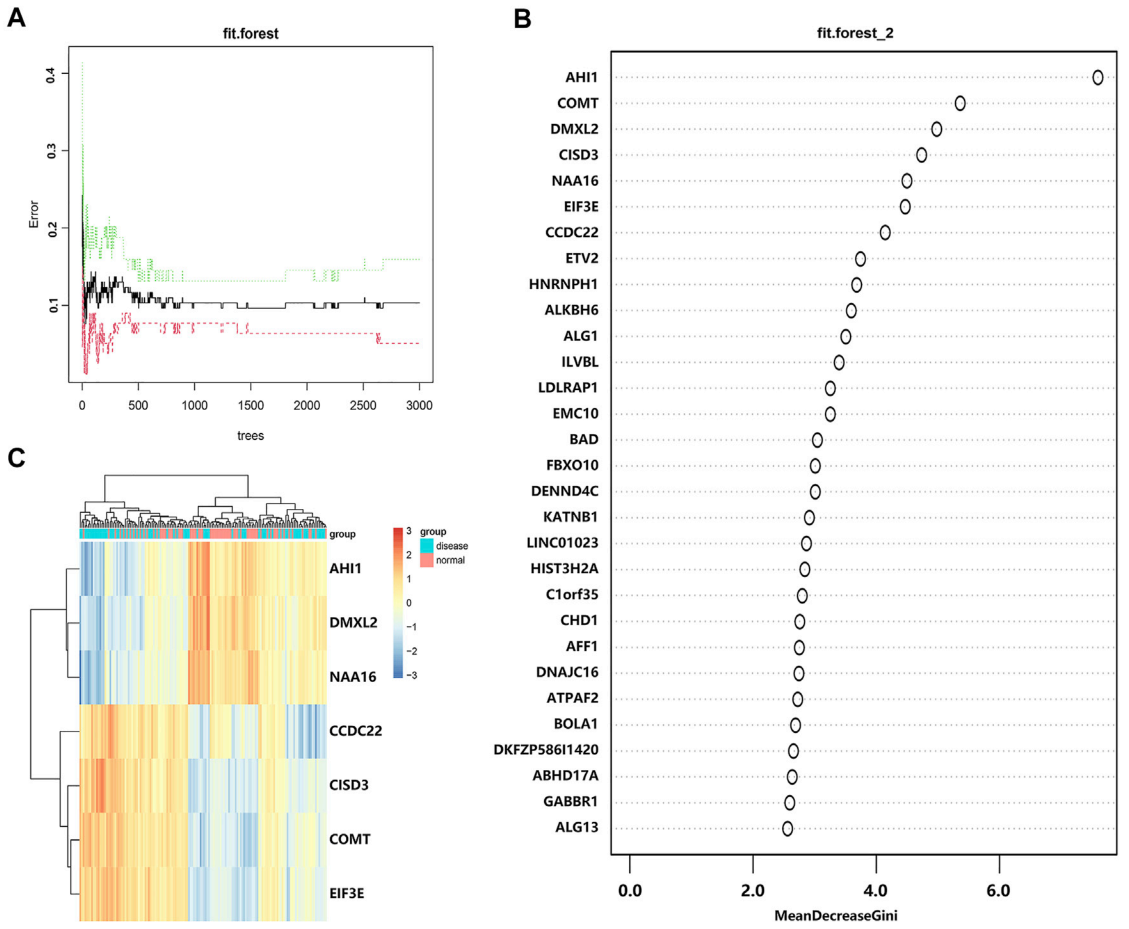


FIGURE 4 | Screening DEGs with the random forest model. **(A)** The relationship between the number of decision tree and the model error. The *x*-axis represents the number of decision trees, and the *y*-axis represents the error rate of the constructed model. When the number of decision trees is nearly 219, the error rate of the constructed model is relatively stable. **(B)** The importance of all variables in the random forest classifier through the Gini coefficient method. The *x*-axis represents the mean decrease of the Gini index, and the *y*-axis represents all variables. **(C)** The heatmap of *k*-means clustering in the GSE6364 dataset. The colors in the graph from red to blue indicate the change from high to low in expression level. On the upper part of the heatmap, the blue band indicates the disease samples and the red band indicates the normal samples.

crucial biomarkers as diagnostic and therapeutic targets of EMs.

Among these seven genes, *COMT*, *NAA16*, *CCDC22*, and *EIF3E* have been reported to be associated with the pathogenesis of EMs. Catechol-*O*-methyltransferase (*COMT*) is highly expressed in the placental, adrenal gland, ovary, and other tissues. The degradative pathways of the catecholamine transmitters can relieve painful uterine contractions (D'Astous-Gauthier et al., 2021). *COMT* polymorphism may contribute to the risk of EMs and adenomyosis (Li et al., 2018) and has a relationship with EM susceptibility (Ji et al., 2017; Zhai et al., 2019). *N*-alpha-acetyltransferase 16 (*NAA16*) is highly enriched in bone marrow, testis, endometrium, and other tissues. It can alter NAT 2 enzyme activity and thus contribute to the susceptibility of EMs (Nakago et al., 2001). Coiled-coil domain containing 22 (*CCDC22*), a membrane-binding protein, is highly enriched in the spleen, lymph node, and other tissues. Studies have demonstrated that there is also a relationship between *CCDC22* polymorphisms and EM susceptibility (de Oliveira Francisco et al., 2017). Eukaryotic translation initiation factor 3 subunit E (*EIF3E*) is highly expressed in the ovary, lymph node, endometrium, and other tissues. Its downregulation may be involved in epithelial–mesenchymal transition (EMT) in EMs, possibly through the preferential translation of snail (an inhibitor of E-cadherin) (Cai et al., 2018) and involved in the development of adenomyosis through activating the TGF-β1 signaling pathway (Cai et al., 2019).

Interestingly, we identified another three important genes (*AHI1*, *DMXL2*, and *CISD3*), which have never been reported to be involved in the pathogenesis of EMs. Abelson helper integration site 1 (*AHI1*) is highly enriched in testis, adrenal gland, brain, prostate, endometrium, and other tissues, which has upregulated expression level in EMs. The *AHI1* protein participates in reactive oxygen species (ROS) production in the

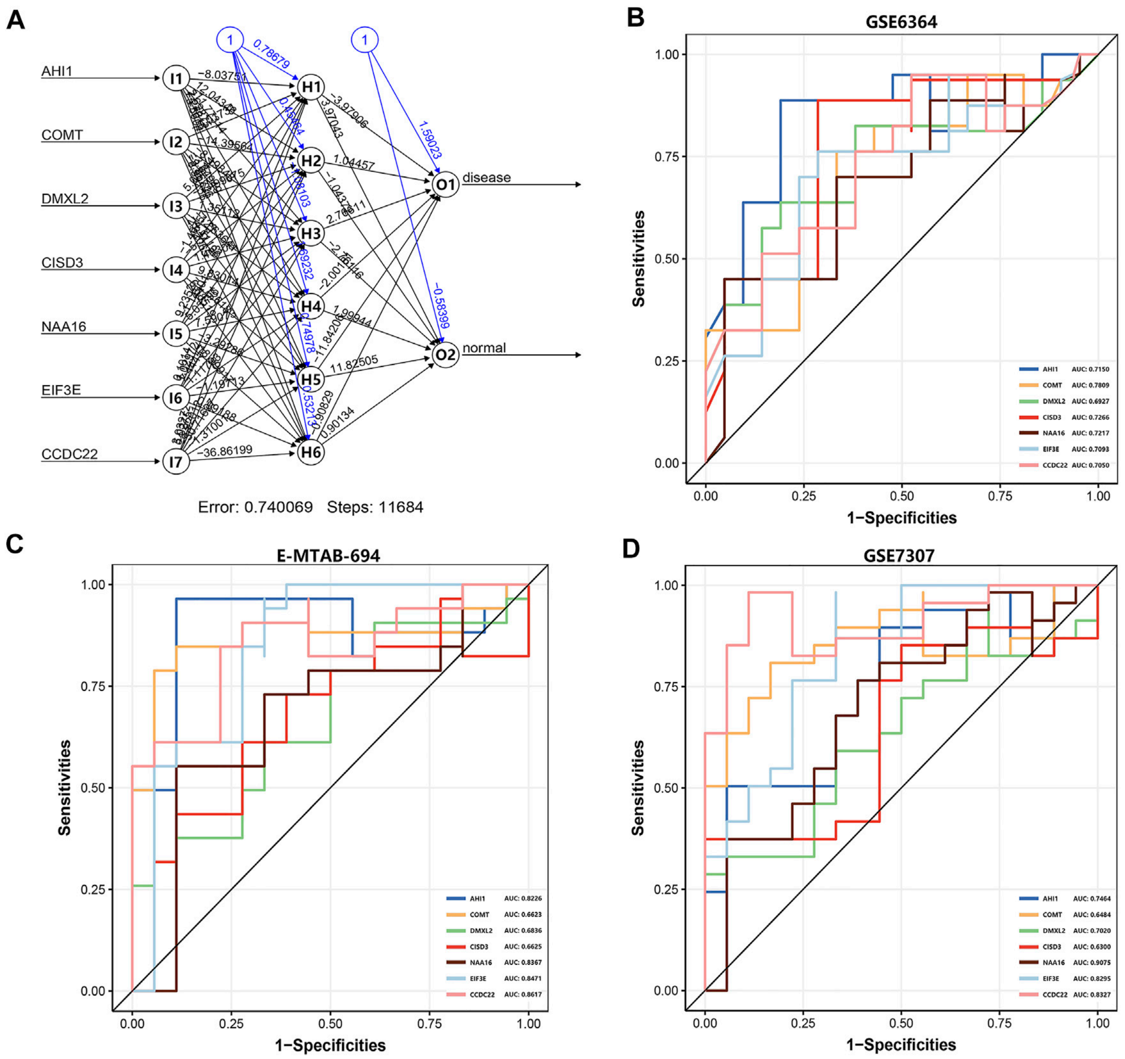


FIGURE 5 | The artificial neural network model and the evaluation of the ROC curve. **(A)** The visualization of the artificial neural network model. **(B)** The evaluation results of the ROC curve in the GSE6364 dataset. **(C)** The verification results of the ROC curve in the E-MTAB-694 dataset. **(D)** The verification results of the ROC curve in the GSE7307 dataset. The *x*-axis and *y*-axis represent specificity and sensitivity, respectively. The AUC value is the area under the ROC curve.

form of protein complexes (Liu et al., 2017). Excessive production of ROS can result in oxidative stress (OS) and overall immune activation and inflammation (Newsholme et al., 2016). OS represents an imbalance between ROS and antioxidants, which may have an essential role in the endometriosis pathogenesis in the peritoneal cavity (Samimi et al., 2019). Hence, the *AHI1* protein may participate in the EMs pathogenesis through multiple processes such as OS and immune and inflammatory response.

Dmx like 2 (*DMXL2*) encodes a protein with 12 WD domains, which has relatively low expression in endometrium tissue and downregulated expression in EMs. The *DMXL2* protein is demonstrated to participate in the regulation of the Notch signaling pathway (Sethi et al., 2010) and acts as a transmembrane protein, which can promote EMT through hyperactivation of the Notch signaling pathway (Faronato et al., 2015). Interestingly, decreased Notch signaling can contribute to impaired decidualization through the downregulation of FOXO1 (a downstream target of Notch signaling) and thus lead to the pathogenesis of EMs (Su et al., 2015). Furthermore, studies indicate that a circRNA with downregulated expression can regulate EMT in EMs via the Notch signaling pathway (Zhang et al., 2019). Therefore, the downregulated expression of *DMXL2* may activate the Notch signaling pathway, contribute to EMT through the interaction with circRNA, and thus lead to the pathogenesis of EMs.

CDGSH iron sulfur domain 3 (*CISD3*) is a member of the CDGSH domain-containing family, whose expression is upregulated in EMs. The *CISD3* protein is redox active and is thought to play an important role in mitochondrial homeostasis (Geldenhuys et al., 2019). Studies indicate that mitochondrial homeostasis can be considered as the therapeutic target for the treatment of EMs via limiting ESC migration and promoting apoptosis (Suliman and Piantadosi, 2016; Zhao et al., 2018). Furthermore, excessive mitochondrial fission can initiate caspase 9-related mitochondrial apoptosis and thus lead to cell death (Fuhrmann and Brüne, 2017; Zhou et al., 2017). Therefore, upregulated expression of *CISD3* may affect mitochondrial homeostasis and thus play an important role in the pathogenesis of EMs.

In this study, based on random forest and artificial neural network algorithm, we established a novel reliable diagnostic model and screened out three important DEGs that have never been reported to be involved in the pathogenesis of EMs. We aimed at the supplement of existing methods and provided an alternative marker panel for further research in the early screening of EMs. However, there are some limitations for this study. Firstly, all samples are only classified as EM (disease) and non-EM (normal) groups, which may affect the final screening results of DEGs. Secondly, the diagnostic model is only verified in two public datasets, which need more samples for verification. Thirdly, we conduct data analysis only at the mRNA level in the tissue samples of EMs, which require further validation at the mRNA and protein levels. In general, our approach has a certain clinical value, which can be beneficial for the early screening of EMs.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**supplementary material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

JS performed data preprocessing and data analysis and wrote the first draft. DS gave advice on the data analysis. RD and LW gave constructive advice for the whole project.

FUNDING

This work was supported by grants from the Science and Technology Innovation Committee of Shenzhen (JCYJ20190806165007495), the Shenzhen Foundation of Science and Technology (JCYJ20210324103606017), and the Guangdong Basic and Applied Basic Research Foundation (2019A1515011693).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.848116/full#supplementary-material

REFERENCES

Bauerle, A., Van Onzenoodt, C., and Ropinski, T. (2021). Net2Vis - A Visual Grammar for Automatically Generating Publication-Tailored CNN Architecture Visualizations. *IEEE Trans. Vis. Comput. Graph.* 27, 2980–2991. doi:10.1109/TVCG.2021.3057483

Beck, M. W. (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *J. Stat. Soft.* 85 (11), 1–20. doi:10.18637/jss.v085.i11

Burney, R. O., and Giudice, L. C. (2012). Pathogenesis and Pathophysiology of Endometriosis. *Fertil. Sterility* 98, 511–519. doi:10.1016/j.fertnstert.2012.06.029

Byeon, H. (2019). Developing a Random forest Classifier for Predicting the Depression and Managing the Health of Caregivers Supporting Patients with Alzheimer's Disease. *Technol. Health Care* 27, 531–544. doi:10.3233/THC-191738

Cai, X., Shen, M., Liu, X., and Guo, S.-W. (2018). Reduced Expression of Eukaryotic Translation Initiation Factor 3 Subunit e and Its Possible Involvement in the Epithelial-Mesenchymal Transition in Endometriosis. *Reprod. Sci.* 25, 102–109. doi:10.1177/1933719117702248

Cai, X., Shen, M., Liu, X., and Nie, J. (2019). The Possible Role of Eukaryotic Translation Initiation Factor 3 Subunit e (eIF3e) in the Epithelial-Mesenchymal Transition in Adenomyosis. *Reprod. Sci.* 26, 377–385. doi:10.1177/1933719118773490

Chen, J., Li, Q., Wang, H., and Deng, M. (2020). A Machine Learning Ensemble Approach Based on Random forest and Radial Basis Function Neural Network for Risk Evaluation of Regional Flood Disaster: A Case Study of the Yangtze River delta, China. *Int. J. Environ. Res. Publ. Health* 17, 49. doi:10.3390/ijerph17010049

Chin, C.-H., Chen, S.-H., Wu, H.-H., Ho, C.-W., Ko, M.-T., and Lin, C.-Y. (2014). cytoHubba: Identifying Hub Objects and Sub-networks from Complex Interactome. *BMC Syst. Biol.* 8 Suppl 4(Suppl 4), S11. doi:10.1186/1752-0509-8-S4-S11

Curchoe, C. L., Flores-Saiffe Farias, A., Mendizabal-Ruiz, G., and Chavez-Badiola, A. (2020). Evaluating Predictive Models in Reproductive Medicine. *Fertil. Sterility* 114, 921–926. doi:10.1016/j.fertnstert.2020.09.159

D'Astous-Gauthier, K., Graham, F., Paradis, L., Des Roches, A., and Bégin, P. (2021). Beta-2 Agonists May Be Superior to Epinephrine to Relieve Severe Anaphylactic Uterine Contractions. *J. Allergy Clin. Immunol. Pract.* 9, 1232–1241. doi:10.1016/j.jaip.2020.10.047

Davis, S., and Meltzer, P. S. (2007). GEOquery: A Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi:10.1093/bioinformatics/btm254

de Oliveira Francisco, D., de Paula Andres, M., Gueuvoghlanian-Silva, B. Y., Podgaec, S., and Fridman, C. (2017). CCDC22 Gene Polymorphism Is Associated with Advanced Stages of Endometriosis in a Sample of Brazilian Women. *J. Assist. Reprod. Genet.* 34, 939–944. doi:10.1007/s10815-017-0936-0

Demchak, B., Hull, T., Reich, M., Liefeld, T., Smoot, M., Ideker, T., et al. (2014). Cytoscape: the Network Visualization Tool for GenomeSpace Workflows. *F1000Res* 3, 151. doi:10.12688/f1000research.4492.2

Drabble, S. J., Long, J., Alele, B., and O'Cathain, A. (2021). Constellations of Pain: a Qualitative Study of the Complexity of Women's Endometriosis-Related Pain. *Br. J. Pain* 15, 345. doi:10.1177/2049463720961413

Faronato, M., Nguyen, V. T. M., Patten, D. K., Lombardo, Y., Steel, J. H., Patel, N., et al. (2015). DMXL2 Drives Epithelial to Mesenchymal Transition in Hormonal Therapy Resistant Breast Cancer through Notch Hyper-Activation. *Oncotarget* 6, 22467–22479. doi:10.18632/oncotarget.4164

Fritsch, S., and Guenther, F. (2016). Neuralnet: Training of Neural Networks. R Package Version 1.33. Available at: https://CRAN.R-project.org/package=neuralnet. 2010-006.

Fuhrmann, D. C., and Brüne, B. (2017). Mitochondrial Composition and Function under the Control of Hypoxia. *Redox Biol.* 12, 208–215. doi:10.1016/j.redox.2017.02.012

Galindo, S. E., Toharia, P., Robles, Ó. D., Ros, E., Pastor, L., and Garrido, J. A. (2020). Simulation, Visualization and Analysis Tools for Pattern Recognition

Assessment with Spiking Neuronal Networks. *Neurocomputing* 400, 309–321. doi:10.1016/j.neucom.2020.02.114

Geldenhuys, W. J., Skolik, R., Konkle, M. E., Menze, M. A., Long, T. E., and Robart, A. R. (2019). Binding of Thiazolidinediones to the Endoplasmic Reticulum Protein Nutrient-Deprivation Autophagy Factor-1. *Bioorg. Med. Chem. Lett.* 29, 901–904. doi:10.1016/j.bmcl.2019.01.041

Gómez-Rubio, V. (2017). ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *J. Stat. Softw.* 77 (2), 678–679. doi:10.18637/jss.v077.b02

Ji, F., Yang, X., He, Y., Wang, H., Aili, A., and Ding, Y. (2017). Aberrant Endometrial DNA Methylome of Homeobox A10 and Catechol-O-Methyltransferase in Endometriosis. *J. Assist. Reprod. Genet.* 34, 409–415. doi:10.1007/s10815-016-0862-6

Khan, M. T., Kaushik, A. C., Ji, L., Malik, S. I., Ali, S., and Wei, D.-Q. (2019). Artificial Neural Networks for Prediction of Tuberculosis Disease. *Front. Microbiol.* 10, 395. doi:10.3389/fmicb.2019.00395

Kong, Y., and Yu, T. (2018). A Deep Neural Network Model Using Random Forest to Extract Feature Representation for Gene Expression Data Classification. *Sci. Rep.* 8, 16477. doi:10.1038/s41598-018-34833-6

Kugunavar, S., and Prabhakar, C. J. (2021). Convolutional Neural Networks for the Diagnosis and Prognosis of the Coronavirus Disease Pandemic. *Vis. Comput. Ind. Biomed. Art* 4, 12. doi:10.1186/s42492-021-00078-w

Li, Y.-w., Wang, C.-x., Chen, J.-s., Chen, L., Zhang, X.-q., and Hu, Y. (2018). Catechol-O-methyltransferase 158G/A Polymorphism and Endometriosis/adenomyosis Susceptibility: A Meta-Analysis in the Chinese Population. *J. Can. Res. Ther.* 14, 980. doi:10.4103/0973-1482.188439

Li, H., Lai, L., and Shen, J. (2020). Development of a Susceptibility Gene Based Novel Predictive Model for the Diagnosis of Ulcerative Colitis Using Random forest and Artificial Neural Network. *Aging* 12, 20471–20482. doi:10.18632/aging.103861

Liaw, A., Yan, J., Li, W., Han, L., Schroff, F., Criminisi, A., et al. (2014). *Package "randomForest"*. R news.

Liu, X., Rothe, K., Yen, R., Fruhstorfer, C., Maetzig, T., Chen, M., et al. (2017). A Novel AHI-1-BCR-ABL-DNM2 Complex Regulates Leukemic Properties of Primitive CML Cells through Enhanced Cellular Endocytosis and ROS-Mediated Autophagy. *Leukemia* 31, 2376–2387. doi:10.1038/leu.2017.108

Nakago, S., Hadfield, R. M., Zondervan, K. T., Mardon, H., Manek, S., Weeks, D. E., et al. (2001). Association between Endometriosis and N-Acetyl Transferase 2 Polymorphisms in a UK Population. *Mol. Hum. Reprod.* 7, 1079–1083. doi:10.1093/molehr/7.11.1079

Newsholme, P., Cruzat, V. F., Keane, K. N., Carlessi, R., and De Bittencourt, P. I. H. (2016). Molecular Mechanisms of ROS Production and Oxidative Stress in Diabetes. *Biochem. J.* 473, 4527–4550. doi:10.1042/BCJ20160503C

Parasar, P., Ozcan, P., and Terry, K. L. (2017). Endometriosis: Epidemiology, Diagnosis and Clinical Management. *Curr. Obstet. Gynecol. Rep.* 6, 34–41. doi:10.1007/s13669-017-0187-1

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies. *Nucleic Acids Res.* 43, e47. doi:10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011). pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinformatics* 12, 77. doi:10.1186/1471-2105-12-77

Samimi, M., Pourhanifeh, M. H., Mehdizadehkashi, A., Eftekhar, T., and Asemi, Z. (2019). The Role of Inflammation, Oxidative Stress, Angiogenesis, and Apoptosis in the Pathophysiology of Endometriosis: Basic Science and New Insights Based on Gene Expression. *J. Cel. Physiol.* 234, 19384–19392. doi:10.1002/jcp.28666

Schonlau, M., and Zou, R. Y. (2020). The Random forest Algorithm for Statistical Learning. *Stat. J.* 20, 3–29. doi:10.1177/1536867X20909688

Sethi, N., Yan, Y., Quek, D., Schupbach, T., and Kang, Y. (2010). Rabconnectin-3 Is a Functional Regulator of Mammalian Notch Signaling. *J. Biol. Chem.* 285, 34757–34764. doi:10.1074/jbc.M110.158634

Shaia, K. L., Acharya, C. R., Smeltzer, S., Lyerly, H. K., and Acharya, K. S. (2019). Non-invasive Diagnosis of Endometriosis: Using Machine Learning Instead of the Operating Room. *Fertil. Sterility* 112, e80. doi:10.1016/j.fertnstert.2019.07.331

Su, R.-W., Strug, M. R., Joshi, N. R., Jeong, J.-W., Miele, L., Lessey, B. A., et al. (2015). Decreased Notch Pathway Signaling in the Endometrium of Women with Endometriosis Impairs Decidualization. *J. Clin. Endocrinol. Metab.* 100, E433–E442. doi:10.1210/jc.2014-3720

Suliman, H. B., and Piantadosi, C. A. (2016). Mitochondrial Quality Control as a Therapeutic Target. *Pharmacol. Rev.* 68, 20–48. doi:10.1124/pr.115.011502

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Taylor, H. S., Adamson, G. D., Diamond, M. P., Goldstein, S. R., Horne, A. W., Missmer, S. A., et al. (2018). An Evidence-Based Approach to Assessing Surgical versus Clinical Diagnosis of Symptomatic Endometriosis. *Int. J. Gynecol. Obstet.* 142, 131–142. doi:10.1002/ijgo.12521

Tian, Y., Yang, J., Lan, M., and Zou, T. (2020). Construction and Analysis of a Joint Diagnosis Model of Random forest and Artificial Neural Network for Heart Failure. *Aging* 12, 26221–26235. doi:10.18632/aging.202405

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., et al. (2021). clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data. *Innovation* 2, 100141. doi:10.1016/j.xinn.2021.100141

Xie, N.-N., Wang, F.-F., Zhou, J., Liu, C., and Qu, F. (2020). Establishment and Analysis of a Combined Diagnostic Model of Polycystic Ovary Syndrome with Random Forest and Artificial Neural Network. *Biomed. Res. Int.* 2020, 1–13. doi:10.1155/2020/2613091

Yigit, A., and Isik, Z. (2018). "Application of Artificial Neural Networks in Dementia and Alzheimer's Diagnosis," in 26th IEEE Signal Processing and Communications Applications Conference, SIU 2018. Izmir, Turkey: May 2–5, 2018. doi:10.1109/SIU.2018.8404447

Yoo, S. M., Choi, J. H., Lee, S. Y., and Yoo, N. C. (2009). Applications of DNA Microarray in Disease Diagnostics. *J. Microbiol. Biotechnol.* 19, 635. doi:10.4014/jmb.0803.226

Zhai, J., Jiang, L., Wen, A., Jia, J., Zhu, L., and Fan, B. (2019). Analysis of the Relationship between COMT Polymorphisms and Endometriosis Susceptibility. *Med. (United States* 98, e13933. doi:10.1097/MD.0000000000013933

Zhang, M., Wang, S., Tang, L., Wang, X., Zhang, T., Xia, X., et al. (2019). Downregulated Circular RNA Hsa_circ_0067301 Regulates Epithelial-Mesenchymal Transition in Endometriosis via the miR-141/Notch Signaling Pathway. *Biochem. Biophys. Res. Commun.* 514, 71–77. doi:10.1016/j.bbrc.2019.04.109

Zhao, Q., Ye, M., Yang, W., Wang, M., Li, M., Gu, C., et al. (2018). Effect of Mst1 on Endometriosis Apoptosis and Migration: Role of Drp1-Related Mitochondrial Fission and Parkin-Required Mitophagy. *Cell. Physiol. Biochem.* 45, 1172–1190. doi:10.1159/000487450

Zhou, H., Zhu, P., Guo, J., Hu, N., Wang, S., Li, D., et al. (2017). Ripk3 Induces Mitochondrial Apoptosis via Inhibition of FUNDC1 Mitophagy in Cardiac IR Injury. *Redox Biol.* 13, 498–507. doi:10.1016/j.redox.2017.07.007

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Genetics

ORIGINAL RESEARCH
published: 14 March 2022
doi: 10.3389/fgene.2022.822666

Explanation-Driven Deep Learning Model for Prediction of Brain Tumour Status Using MRI Image Data

*Loveleen Gaur[1], Mohan Bhandari[2], Tanvi Razdan[1], Saurav Mallik[3] and Zhongming Zhao[3,4]**

[1]Amity International Business School, Amity University, Noida, India, [2]Nepal College of Information Technology, Lalitpur, Nepal, [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States, [4]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States

Cancer research has seen explosive development exploring deep learning (DL) techniques for analysing magnetic resonance imaging (MRI) images for predicting brain tumours. We have observed a substantial gap in explanation, interpretability, and high accuracy for DL models. Consequently, we propose an explanation-driven DL model by utilising a convolutional neural network (CNN), local interpretable model-agnostic explanation (LIME), and Shapley additive explanation (SHAP) for the prediction of discrete subtypes of brain tumours (meningioma, glioma, and pituitary) using an MRI image dataset. Unlike previous models, our model used a dual-input CNN approach to prevail over the classification challenge with images of inferior quality in terms of noise and metal artifacts by adding Gaussian noise. Our CNN training results reveal 94.64% accuracy as compared to other state-of-the-art methods. We used SHAP to ensure consistency and local accuracy for interpretation as Shapley values examine all future predictions applying all possible combinations of inputs. In contrast, LIME constructs sparse linear models around each prediction to illustrate how the model operates in the immediate area. Our emphasis for this study is interpretability and high accuracy, which is critical for realising disparities in predictive performance, helpful in developing trust, and essential in integration into clinical practice. The proposed method has a vast clinical application that could potentially be used for mass screening in resource-constraint countries.

Keywords: LIME, SHAP, XAI, brain tumor, MRI

OPEN ACCESS

Edited by:
Alfonso Maurizio Urso,
Institute for High Performance Computing and Networking (ICAR)—Italian National Research Council (CNR), Italy

Reviewed by:
Andrea Tangherloni,
University of Bergamo, Italy
Leonardo Rundo,
University of Salerno, Italy

****Correspondence:***
Zhongming Zhao
Zhongming.Zhao@uth.tmc.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *26 November 2021*
Accepted: *17 February 2022*
Published: *14 March 2022*

Citation:
Gaur L, Bhandari M, Razdan T, Mallik S and Zhao Z (2022) Explanation-Driven Deep Learning Model for Prediction of Brain Tumour Status Using MRI Image Data.
Front. Genet. 13:822666.
doi: 10.3389/fgene.2022.822666

1 INTRODUCTION

According to the world health organization (WHO) world cancer report (2020), cancer is amongst the leading death-causing diseases, ranked second (after cardiovascular disease), accounting for nearly 10 million deaths in 2020 (Sung et al., 2021). Compared to other diagnoses, cancer screening is a different and more complicated public health approach that needs extra resources, infrastructure, and coordination. The WHO recommends the implementation of screening programs when the following conditions are fulfilled (Sung et al., 2021):

1. The efficiency of tool/model/software has been demonstrated

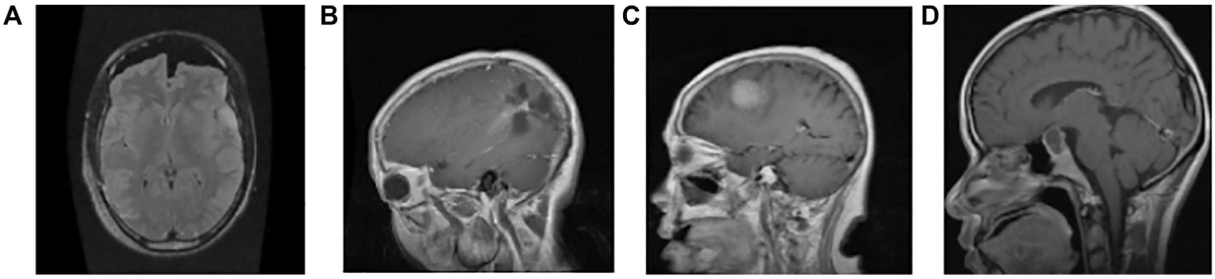


FIGURE 1 | Sample image data of different types of tumours. **(A)** Normal: the intensity of the parenchyma in the brain without any tumour is normal. The ventricular system and cisternal spaces are supposed to be in good working order. There is always no evidence of an intracranial space-occupying lesion (Gaillard, 2021). **(B)** Glioma tumour: gliomas have thick, irregularly enhancing borders of the focal necrotic core with a haemorrhagic component. They are surrounded by vasogenic-type oedema, containing malignant cell infiltration. Intratumoural haemorrhage happens rarely (less than 2%) (Frank, 2021) **(C)** Meningioma tumour: meningiomas are extra-axial tumours arising from meningocytes or arachnoid cap cells of meninges and can be found where meninges exist, as well as in some sites where only rest cells are thought to exist (Gaillard and Rasuli, 2021) **(D)** Pituitary tumour: for pituitary adenomas, minor intra-pituitary lesions appear differently than larger lesions that spread into the suprasellar region and pose various surgical and diagnostic issues. Based on tumour aspects, overall signal qualities can vary (Weerakkody and Gaillard, 2021).

2. Sufficient resources and facilities to confirm diagnoses and treatments are available
3. The prevalence of the disease is extreme enough to justify the screening

The total prevalence of all central nervous system tumours is 3.9 per 100,000 persons worldwide; the incidence differs with age, gender, race, and region and is extremely frequent in Northern Europe, followed by Australia, the United States, and Canada. Meningioma is the most common one, accounting for 36.8% of all tumours; glioma is the most widespread malignant tumour, accounting for 75% of central nervous system malignant tumours, with a total incidence of six cases per 100,000 people per year. MRI is presently the ideal method for early detection of human brain tumours as it is non-invasive (Spatharou et al., 2021). However, the interpretation of MRI is predominantly centred on the opinions of radiologists.

The advent of convolution neural network (CNN)-based deep learning (DL) provides the basis for imaging-based artificial intelligence (AI) solutions. DL-guided solutions intend to supplement clinical decision making. There are several motives why the proposed architecture is a CNN-based DL architecture. First, it is observed that CNN-based DL is extremely good at lowering the threshold of parameters while maintaining model quality. Second, it does not require human feature engineering because it can automatically extract features from an image. Third, the literature supports the CNN-based DL model by several researchers and that it has achieved good image classification and recognition accuracy. However, it is crucial to observe that very few researchers have applied local interpretable model-agnostic explanation (LIME) and Shapley additive explanation (SHAP) along with CNN. Researchers demonstrated the immense potential of imaging tools to mitigate the heavy burden on medical experts (Wojciech et al., 2017). It further allows devoting additional help in patient care, reducing burnout, and shrinking overall medical costs for patients (Dave et al., 2020). Working on the detection system, Gupta et al. (2016) applied DL algorithms, Resnet50, to distinguish COVID-19 from X-rays to achieve a fully autonomous and speedier diagnosis. With an average COVID-19 detection time of roughly 2.5 s and an average accuracy of 0.97, the authors aimed to minimise the run time to about 2.5 s. Kollias et al. (2018) introduced different performance indicators such as precision, responsiveness, specificity, precision, F1 value, and DL. The results showed a standard accuracy of 92.93% and sensitivity of 94.79% to provide robust identification and detection of COVID-19 in the chest X-ray dataset. In one of the research (Ke et al., 2019), the deep neural network correlation learning mechanism for CT brain tumour detection used palettes of CNN architecture to adjust them to the best possible detection result of ANN. The AISA framework for MRI data analysis demonstrated its application to brain scan data by deriving independent subspaces and extracting texture features. Then, dimensionality is reduced using t-SNE embedding for discriminative classification. Finally, the KNN classification is applied. Despite the immense popularity of DL models in clinical decision making, the lack of interpretability and transparency by algorithm-driven decisions remains the biggest challenge, particularly in medical settings. Although, many researchers (Richard et al., 2020; Zucco et al., 2018) observed various impediments in developing XAI-based clinical decision support systems (CDSS) due to the non-availability of any universal notion of explainability. Our study proposes an explanation-driven DL-based model to predict distinctive subtypes of brain tumours (meningioma, glioma, and pituitary) using an MRI image dataset. We also implemented LIME and Shapley additive explanations to create more transparency in the models while keeping intact a high performance rate. Our study will help the users (medical professionals, clinicians, etc.) in comprehending and efficiently managing the ever-increasing number of trustable and reliable AI partners (Sharma et al., 2020).

Compared to previous models, our model used a dual-input CNN approach to prevail over the classification challenge with inferior-quality images and an accuracy of 94.64% compared to other state-of-the-art models. Previous studies lack explanation, and thus, we used Explainable AI (XAI) algorithms such as LIME and SHAP, which is the differentiating element of this study. We used SHAP to ensure consistency and local accuracy for interpretation as Shapley values

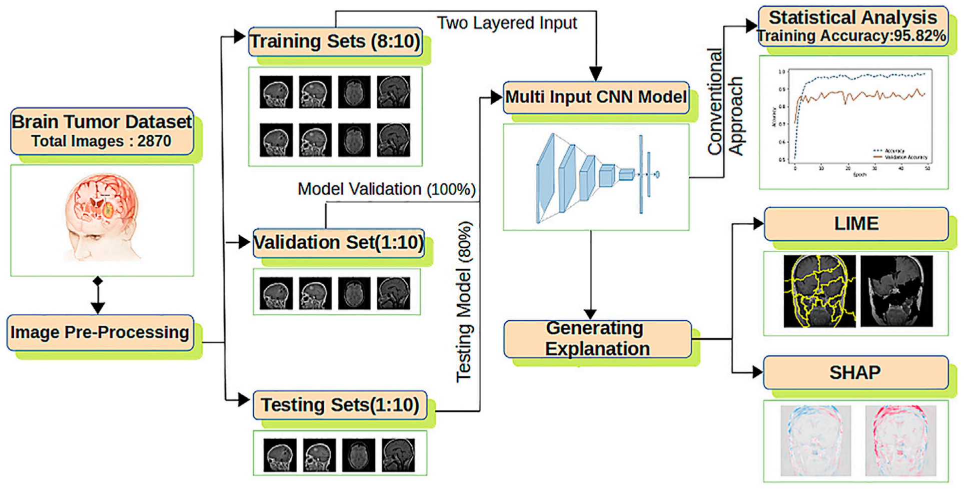


FIGURE 2 | The proposed explanation-driven DL model for prediction of brain tumour status using MRI image data: 2870 MRI images are pre-processed and divided into training, validation, and test sets. Two copies of datasets are fed into a multi-input CNN model to find the training, validation, and test accuracy. The same CNN model was further imposed on LIME and SHAP.

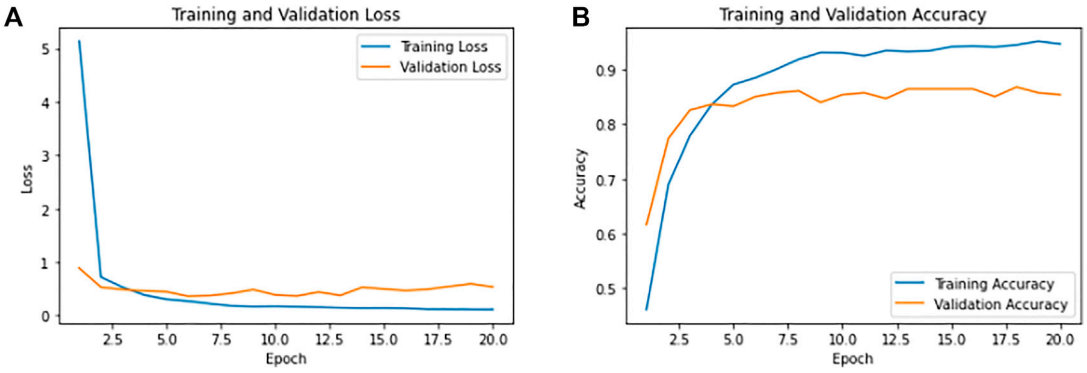


FIGURE 3 | Training and validation results of CNN. **(A)** Shows the overall training loss as 0.1149 and validation loss as 0.53. **(B)** Shows the overall training accuracy as 94.64% and validation accuracy of 85.37%.

examine all potential predictions using all possible combinations of inputs. Conversely, LIME constructs sparse linear models around each prediction to describe how the model operates in the immediate area.

The deep neural network correlation learning mechanism for computed tomography (CT) brain tumour detection used palettes of CNN architecture to adjust them to the best possible detection result of DL. Though the previously suggested models have higher accuracy, they lack explainability, interpretability, and transparency (Abdalla and Esmail, 2018; Khairandish et al., 2021). The proposed model used XAI algorithms such as LIME (Vedaldi and Soatto, 2008) and SHAP as detailed in Algorithm 2.

The contributions in this study are summarised in what follows:

1. We aimed to create an explanation-driven multi-input DL model where SHAP and LIME are used for an in-depth description of results. One set of two input datasets is fed to the convolution layer and one to the fully connected layer.
2. We have achieved high accuracy of (94.64%) brain MRI images compared to other state-of-the-art models.

2 METHODS

2.1 Datasets

In this study, we used the publicly available MRI images (Bhuvaji, 2020). The datasets are annotated into three categories of tumours: glioma tumour, meningioma tumour, and pituitary tumour, along with the normal image. Out of 2,870 total images, 2,296 images of distinct types are used as training sets and the remaining as test sets.

TABLE 1 | K-fold cross-validation results.

Fold	Final validation loss	Final validation accuracy (%)
1	0.011 44	99.5
2	0.017 06	98.47
3	0.021 52	99.13
4	0.009 88	99.34
5	0.005 54	100
6	0.012 98	99.13
7	0.008 74	99.78
8	0.005 33	99.78
9	0.010 18	99.34
10	0.008 8	99.56

2.1.1 Data Pre-Processing

All 512 × 512 × 3 images are resized to 150 × 150 × 3. The images are rearranged for faster convergence and preventing the CNN model from learning the training order. For better classification results, we have introduced Gaussian noise as it improves the learning for DL (Neelakantan et al., 2015) with mean = 0 and standard deviation $10^{0.5}$. **Figure 1** shows a single instance among the categories of tumours from the dataset.

2.2 Proposed Framework

The overall architecture of the model used is shown in **Figure 2** composed of feature extraction, a CNN model, statistical performance measures, and explanation extraction frameworks.

For improved accuracy, two copies of the dataset are fed to the CNN model having an output layer of size 1 × 4 and six hidden layers (Yu et al., 2017). Adam optimiser with its default parameters is applied with the rectified linear unit (ReLU) and softmax as the activation function. The final CNN model is used for statistical accuracy measurement, LIME and SHAP. For LIME explanations, perturbation is calculated, whereas for SHAP, a gradient explainer is applied. The whole process is formalized in Algorithm 1.

Algorithm 1. Explanation-driven multi-input DL model for prediction of brain tumour.

```
Input: MRI Dataset(Break down in ratio 8:1:1) with size 150x150x3
Epoch: 20
Optimizer: Adam
Kernel Size: 3 x 3
Dropout: 0.2
Filter : Conv2D
for For every iteration in dual input CNN Model do
    Input one set of dataset to convolution layer
    Input one set of dataset to fully connected layer
    Calculate loss, accuracy, validation loss, validation accuracy
end for
Implement XAI: Implement LIME and SHAP for the model
```

For the classification task in the proposed explainable model, a CNN with dual-input architecture is used. The CNN is imposed with ReLU as activation in all hidden layers. Compared with the input value and zero value, ReLU is simple to calculate. Furthermore, ReLU has a derivative of either 0 or 1 based on positive or negative input. This feature of ReLU is essential in comparing explainable modules such as LIME and SHAP. Adam optimiser with its default parameter (Kingma and Ba, 2015) is used along with sparse categorical cross entropy; the kernel size is set to 3 × 3.

3 RESULTS

Following the classification process, the performance of CNN models is evaluated based on accuracy and the number of wrong predictions. The curves for the conventional results of CNN are presented in **Figure 3**.

3.1 CNN

The model was iterated for 20 epochs, and during callback in CNN modules, we had monitored the loss with min mode and patience level of three to cross the over-fitting. Achieving the training accuracy of 94.64% and overall test accuracy of 85.37%, the model has 26 wrong predictions with 0.1149 as training loss and 0.53 as validation loss.

Furthermore, to estimate the performance of the CNN model on the configured dataset, K-fold cross validation is performed with K = 10 non-overlapping folds for 20 epochs with a batch size of 128. The test and train sets were split in the ratio of 1:4. The final validation result of the cross fold is shown in **Table 1**. The proposed model has achieved almost 100% training accuracy during cross validation.

Table 2 shows the confusion matrix for 287 test images. A total of 7 normal images out of 46, 14 glioma images out of 84, 12 meningioma images out of 77, and 3 pituitary images out of 80 were misclassified.

To validate our model statistically, we performed McNemar's test (Smith et al., 2020). For labels of test data and labels of model prediction under test data, McNemar's test gave a chi-squared value of 42.022 and p value 9.02 x e^{-11}. We can reject the null-hypothesis that both labels perform equally well on the test set, since the p value is smaller than $\alpha = 0.005$.

3.2 SHAP

For each pixel on a predicted image, the scores show its contribution and can be used to explain tumour classification tasks. The Shapley values correspond to each feature for different categories of the tumour according to Algorithm 2.

Algorithm 2. Algorithm to calculate the Shapley values.

```
Input: Number of iterations M, instance of interest x, feature index j, data matrix X, and Dual input CNN model
for Every Iteration 1 ..... M do
    Draw random instance z from the data matrix X
    Choose a random permutation o of the feature values
    Order instance x: ($x_0$ ...., $x_j$, ........... , $x_p$)
    Order instance z: ($z_0$ ...., $z_j$, ........... , $z_p$)
    Construct two new instances:
        with j:$x_{-j}$ = ( $x_0$ ..., $x_j$,$z_{j+1}$, ........ $z_p$)
        without j:$x_{+j}$ = ( $x_0$ ..., $x_j$,$z_{j+1}$, ........ $z_p$)
    Compute Marginal Distribution:
        $\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$    (1)
end for
Compute shapely values:
    $\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$    (2)
```

TABLE 2 | Confusion matrix for the CNN.

		Actual value			
		Normal	Glioma	Meningioma	Pituitary
Predicted values	Normal	37	8	1	0
	Glioma	7	70	5	2
	Meningioma	0	12	65	0
	Pituitary	0	3	0	77

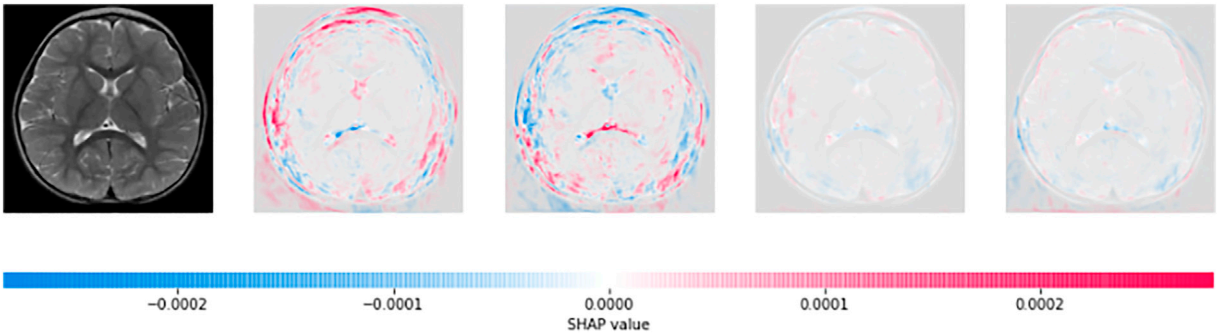


FIGURE 4 | On the basis of Shapley values, we can say that the MRI image is normal.

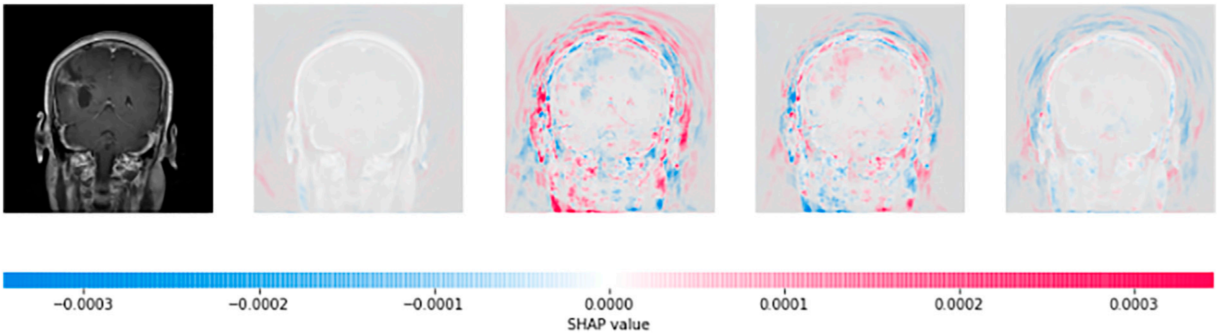


FIGURE 5 | On the basis of Shapley values, we can say that the MRI image holds meningioma tumour.

The CNN model with mathematical behaviour is complicated to interpret directly. Thus, the effect of individual input features on the model's output is clearly explained using SHAP and shown in **Figures 4**, **5**. Positive SHAP values that raise the likelihood of the class are represented by red pixels. In contrast, negative SHAP values that lower the probability of the class are represented by blue pixels. **Figure 4** and **Figure 5** are test images. In contrast, the rest of the figures indicate the normal image and three other categories of tumour: glioma, meningioma, and pituitary tumours in successive order.

3.3 LIME

A total of 150 perturbations are used. Random ones and zeros are produced and formed into a matrix, with perturbations as rows and superpixels as columns. A superpixel is ON if it is 1, and it is OFF if it is 0. The length of the displayed vector represents the number of superpixels in the image. The test image is perturbed based on the perturbation vector and predefined superpixels (Vedaldi and Soatto, 2008). The final perturbed image is shown in **Figure 6C** for normal test image under consideration and in **Figure 7C** for test image under consideration with meningioma tumour, which shows the portion of the image having a major role for classification.

The CNN model is utilised to generate the explanation using LIME. **Figure 6A**is a normal image, and **Figure 7A** is under the meningioma category. The classification produces a vector of 2,870 probabilities for each category accessible in the CNN model. The quick-shift segmentation method is used to create superpixels. 22 superpixels are generated for **Figure 6A** and

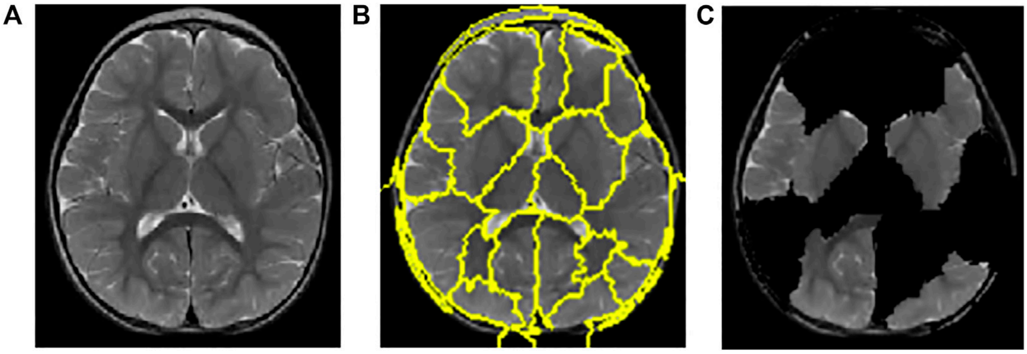


FIGURE 6 | Interpretations generated by LIME for a normal image. **(A)** Sample of the normal image from the test image. **(B)** Superpixels generated from a sample of the normal image from test image quick-shift segmentation to create perturbations. **(C)** Final perturbed image for the normal image.

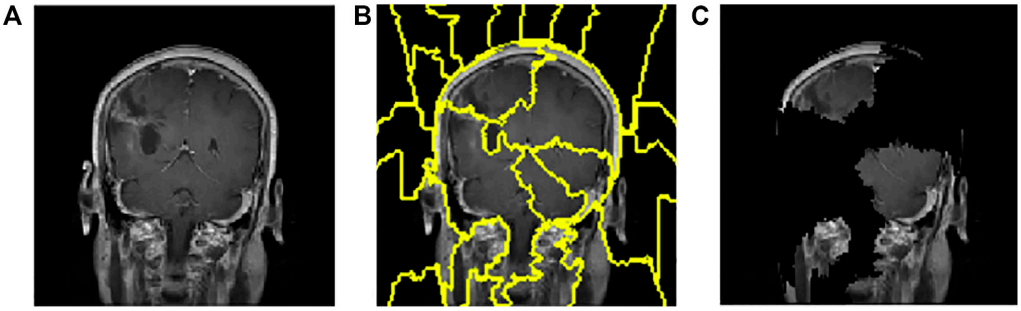


FIGURE 7 | Interpretations generated by LIME for meningioma tumour. **(A)** Sample of meningioma tumour from the test image. **(B)** Superpixels generated from quick-shift segmentation to create perturbations. **(C)** Final perturbed image showing meningioma tumour.

TABLE 3 | Brain tumour detection using traditional ML methods.

Authors	Algorithm	Dataset	Accuracy (%)	XAI
Martinez et al. (2020)	Random Forest	BraTs Dataset	76	No
Minz and Mahobiya (2017)	Adaboost Classifier	BraTs Dataset	89.90	No
Abdalla and Esmail (2018)	Back-Propagation Network	MRI Images	99	No
Asodekar and Gore (2019)	Random Forest	BraTs Dataset	81.90	No
Asodekar and Gore (2019)	SVM	BraTs Dataset	78.57	No
Proposed model	Dual-Input CNN	MRI Images	94.64	Yes

shown in **Figure 6B**, and 24 superpixels are calculated for **Figure 7A** and shown in **Figure 7B**.

4 DISCUSSION

4.1 Comparison of the Proposed Feature Extraction Methods Using Traditional Machine learning (ML) Methods

We compare the proposed feature extraction methods to traditional ML methods. The comparative results are presented in **Table 3**. Minz and Mahobiya (2017) pre-processed the MICCAI BraTS dataset to eliminate noise and employed the GLCM (gray-level co-occurrence matrix) for feature extraction and classification boosting (Adaboost). An MRI was used to extract 22 characteristics. The Adaboost classifier is utilised for classification, and the suggested system achieves a maximum accuracy of 89.90%. Abdalla and Esmail (2018) executed a computer-aided detection system after collecting the MRI images. They processed the image before implementing the back-propagation algorithm and extracted the features using Haralick's features based on the spatial gray-level dependency matrix (SGLD). The results were 99%, but the study could not focus on the explainable section in the training images. A comparative study between support vector machine (SVM)

TABLE 4 | Brain tumour detection using other state-of-the-art models.

Authors	Algorithm	Dataset	Accuracy (%)	XAI
Shahzadi et al. (2018)	CNN with LSTM	MRI Images	84	No
Hemanth et al. (2019)	CNN	MRI Images	91	No
Avsar and Salcin (2019)	R-CNN	MRI Images	91.66	No
Ranjbarzadeh et al. (2021)	C-CNN	BraTs Dataset	92.03	No
Khairandish et al. (2021)	CNN–SVM	MRI Images	98.49	No
Proposed model	Dual-Input CNN	MRI Images	94.69	Yes

and random forest (RF) classified benign and malignant tumours. First, the brain tumour's region of interest was determined for feature extraction, and then, features were calculated. Shape characteristics were obtained and utilised to classify benign and malignant tumours. According to the authors, RF (81.90%) outperformed the SVM (78.57%). By combining principal component analysis (PCA), KSVM, and GRB kernels, Arora and Ratan (2021) established a unique technique for categorisation of MRI brain images using discrete wavelet transform (DWT). The experiment was carried out with four different kernels. The findings demonstrate that combining DWT, PCA, KSVM, and the GRB kernel yields the highest accuracy compared to other methodologies. The results show that the time it takes to classify a segmented picture significantly decreases, which might be a watershed moment in the medical profession for tumour diagnosis. Martinez et al. (2020) worked on the FLAIR images on the BRATS 2015 training dataset; it is used to restructure and increase data attributes that lead to a pixel-based classifier. The U-net suggested method performs a semantic segmentation with a precision of 76%, which increases by 23% compared to the random forest classifier with synthetic minority oversampling technique (SMOTE) class balancing algorithm.

4.2 Comparison of the Proposed Method With the Other State-of-the-Art Methods

This section compares our dual-input CNN model with other state-of-the-art models. The results are compared in **Table 4**. After several data-collection and pre-processing steps such as average filtering segmentation, the DL model was implemented by researchers (Hemanth et al., 2019). In comparison to existing approaches such as conditional random field (89%), SVM (84.5%), and genetic algorithm (GA) (83.64%), the research represents overall performance and comparative output on the brain MRI images. In contrast to existing algorithms, the suggested CNN (91%) produces improved results. The TensorFlow library was used to construct a DL method called faster R-CNN in the work of Avsar and Salcin (2019), and the classifier algorithm was trained and tested using a publicly available dataset of 3,064 MRI brain pictures (708 meningiomas, 1,426 gliomas, and 930 pituitary gland tumours) from 233 patients. The quicker RCNN algorithm has been demonstrated to attain 91.66% accuracy, which is exceptional compared to past work on the same dataset. Ranjbarzadeh et al. (2021) proposed a cascaded convolutional neural network (C-ConvNet/C-CNN). A simple but effective cascade, the CNN model, has been suggested to extract local and global characteristics in two methods, with different extraction patches in each. Those patches were chosen to feed the network that their centre was located inside this area after extracting the tumour's predicted location using a sophisticated pre-processing strategy. As a result of removing a high number of insignificant pixels from the picture in the pre-processing stage, the computing time and ability to generate quick predictions for categorising the clinical image are reduced. The results were compared to other algorithms. Still, the CNN model achieved the highest accuracy (92.03%) on the whole Dice score (mean) and the highest precision (97.12%) on the core sensitivity score (mean). Khairandish et al. (2021) made use of a hybrid model of CNN and SVM in phrases of classification, type, and threshold-based segmentation in terms of detection to classify benign and malignant tumours in brain MRI images. This hybrid CNN–SVM is rated as having an overall accuracy of 98.49%. Still, their study does not show evidence for manipulating low-quality images and XAI. Shahzadi et al. (2018) proposed a CNN cascade with a long short-term memory (LSTM) network for classifying 3D brain tumour MRIs into HG and LG glioma. The features from the pre-trained VGG-16 were retrieved and fed into an LSTM network for learning high-level feature representations. The components extracted from VGG-16 had a classification accuracy of 84%, higher than that of those extracted from AlexNet and ResNet, 71%. Isola et al. (2018) investigated conditional adversarial networks as a general-purpose solution for image-to-image translation challenges by using a 1,616 PatchGAN. The PatchGAN 70 × 70 reduces these distortions and improves scores slightly. It is observed that scaling to the full 286 × 286 ImageGAN does not significantly improve the visual quality of the findings and results in a considerably lower FCN-score, indicating that conditional adversarial networks are a promising option for many image-to-image translation tasks, especially those involving highly structured graphical outputs. Milletari et al. (2016) proposed an approach to 3D image segmentation based on a volumetric, fully convolutional neural network. The CNN is trained end-to-end on MRI volumes depicting the prostate and predicts segmentation for the whole volume at once. The training was performed on 50 MRI volumes, and the relative manual ground truth annotation was obtained from the PROMISE2012 challenge dataset. The novel objective function was to optimise during training based on the dice overlap coefficient between the predicted segmentation and the ground truth annotation. Han et al. (2020) proposed an unsupervised medical anomaly detection generative adversarial network (MADGAN). This two-step method uses GAN-based multiple adjacent brain MRI slice reconstruction to detect brain anomalies at various stages on multi-sequence structural MRI. MADGAN can detect anomaly on T1 scans at a very early stage, mild cognitive impairment (MCI), with area under the curve (AUC) 0.727, and anomaly detection (AD) at a late stage with AUC 0.894, while detecting brain metastases on T1c scans with AUC 0.921. On multi-sequence MRI, the model may accurately detect the accumulation of subtle anatomical abnormalities and hyper-intense enhancing lesions, such as (particularly late stage) AD and brain metastases, as the first unsupervised varied disease diagnosis. Baur et al. (2020) presented a novel method towards unsupervised AD in brain MRI by embedding the modelling of healthy anatomy into a CycleGAN-based style-transfer task, which is trained to translate healthy brain MRI images to a simulated distribution with lower entropy and vice versa. By filtering high-frequency, low-amplitude signals from lower entropy samples during training, the resulting model suppresses anomalies in reconstructing the input data at test time. The method outperforms the state-of-the-art method in various measures and can deal with high-resolution data, a current pitfall of autoencoder (AE)-based methods. Castiglioni et al. (2021) concentrated on the issues that must be addressed to create AI applications as clinical decision support systems in a real-world setting. A narrative review with a critical appraisal of publications published between 1989 and 2021 was conducted. According to the study, biomedical and healthcare systems are among the most significant domains for AI applications, with medical imaging being the most suited and promising domain. Clarification of specific challenging points facilitates the development of such systems and their translation to clinical practice. Barragán-Montero et al. (2021) showcased the technological pillars of AI, as well as the state-of-the-art methods and their implementation to medical imaging. This review offered an overview of AI, emphasising medical imaging analysis demonstrating the potential of the state-of-the-art ML and DL algorithms to automate and enhance several aspects of clinical practice.

5 CONCLUSION AND FUTURE DIRECTION

Using an explanation-driven dual-input CNN model for finding if a particular MRI image is subjected to a tumour or not, the proposed study achieved an accuracy of 94.64%. A brain MRI image dataset is used to train and test the proposed CNN model, and the same model was further imposed to SHAP and LIME algorithms for an explanation. Our experiment utilised two dataset copies as input for better feature extraction, one in the convolution layer and another in the fully connected layer. However, any attempt to remove any features decreased the prediction model's overall performance; hence, no augmentation was carried out. The proposed model is a locally interpreted model with a model-agnostic explanation, shapely explained to describe the results for ordinary people more qualitatively.

In future, classification algorithms with higher accuracy and better optimiser can be used and imposed on XAI. For better clinical issues, the research may be replicated and applied to other XAI algorithms such as GradCAM. Furthermore, like the most recent advances on computing capacity, neuroimaging technologies, and digital phenotyping tools (Ressler and Williams, 2020), algorithms to imitate natural occurrences can be used on heterogeneous datasets for medical imaging modalities, electronic health record engines, multi-omics studies, and real-time monitoring (Rundo et al., 2019).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found here at https://www.kaggle.com/sartajbhuvaji/brain-tumor-classification-mri.

ETHICS STATEMENT

Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

Conceptualisation of the research topic and writing of the original draft were carried out by LG, MB, and TR. Resource collection and design of the methodology and code were carried out LG and MB. Project administration was conducted by LG, SM, and ZZ. Result validation was performed by LG, MB, SM, and ZZ. Final draft and revisions were made by SM and ZZ. Finally, the fund was acquired by ZZ.

FUNDING

ZZ was partially supported by the Cancer Prevention and Research Institute of Texas (CPRIT 180734). The funders did not participate in the study design, data analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Abdalla, H. E. M., and Esmail, M. Y. (2018). "Brain Tumor Detection by Using Artificial Neural Network," in 2018 International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), 1–6. doi:10.1109/iccceee.2018.8515763

Arora, P., and Ratan, R. (2021). "Development of a Novel Approach for Classification of Mri Brain Images Using Dwt by Integrating Pca, Ksvm and Grb Kernel," in Proceedings of Second International Conference on Smart Energy and Communication. doi:10.1007/978-981-15-6707-0_13

Asodekar, B., and Gore, S. A. (2019). *Brain Tumor Classification Using Shape Analysis of Mri Images.*

Avşar, E., and Salçin, K. (2019). Detection and Classification of Brain Tumours from Mri Images Using Faster R-Cnn. *Teh. Glas. (Online)* 13, 337–342. doi:10.31803/tg-20190712095507

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., et al. (2021). Artificial Intelligence and Machine Learning for Medical Imaging: A Technology Review. *Physica Med.* 83, 242–256. doi:10.1016/j.ejmp.2021.04.016

Baur, C., Graf, R., Wiestler, B., Albarqouni, S., and Navab, N. (2020). "Steganomaly: Inhibiting Cyclegan Steganography for Unsupervised Anomaly Detection in Brain Mri," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Editors A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, et al. (Cham: Springer International Publishing), 718–727. doi:10.1007/978-3-030-59713-9_69

Bhuvaji, S. (2020). *Brain Tumor Classification-Mri. Vol. 2*. doi:10.34740/kaggle/dsv/1183165

Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., et al. (2021). Ai Applications to Medical Images: From Machine Learning to Deep Learning. *Physica Med.* 83, 9–24. doi:10.1016/j.ejmp.2021.02.006

Dave, D., Naik, H., Singhal, S., and Patel, P. (2020). *Explainable AI Meets Healthcare: A Study on Heart Disease Dataset. Vol. abs/2011.03195*.

Di Muzio, B., and Gaillard, F. (2021). *Normal Brain Mri: Radiology Case*. doi:10.53347/rID-42777

Frank, G. (2021). *Glioblastoma: Radiology Reference Article*. doi:10.53347/rID-4910

Gaillard, F., and Rasuli, B. (2021). *Meningioma*. doi:10.53347/rID-1659

Gupta, M., Rao, P., and Rajagopalan, V. (2016). "Brain Tumor Detection in Conventional Mr Images Based on Statistical Texture and Morphological Features," in 2016 International Conference on Information Technology (ICIT), 129–133. doi:10.1109/icit.2016.037

Han, C., Rundo, L., Murao, K., Noguchi, T., Shimahara, Y., Milacski, Z., et al. (2020). Madgan: Unsupervised Medical Anomaly Detection gan Using Multiple Adjacent Brain Mri Slice Reconstruction. *BMC Bioinformatics*. In Press.

Hemanth, G., Janardhan, M., and Sujihelen, L. (2019). "Design and Implementing Brain Tumor Detection Using Machine Learning Approach," in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 1289–1294. doi:10.1109/ICOEI.2019.8862553

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2018). *Image-to-image Translation with Conditional Adversarial Networks*.

Ke, Q., Zhang, J., Wei, W., Damasevicius, R., and Wozniak, M. (2019). Adaptive Independent Subspace Analysis of Brain Magnetic Resonance Imaging Data. *IEEE Access* 7, 12252–12261. doi:10.1109/ACCESS.2019.2893496

Khairandish, M. O., Sharma, M., Jain, V., Chatterjee, J. M., and Jhanjhi, N. Z. (2021). A Hybrid Cnn-Svm Threshold Segmentation Approach for Tumor Detection and Classification of Mri Brain Images. *Irbm*. doi:10.1016/j.irbm.2021.06.003

Kingma, D. P., and Ba, J. (2015). "Adam: A Method for Stochastic Optimization," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. Editors Y. Bengio and Y. LeCun.

Kollias, D., Tagaris, A., Stafylopatis, A., Kollias, S., and Tagaris, G. (2018). Deep Neural Architectures for Prediction in Healthcare. *Complex Intell. Syst.* 4, 119–131. doi:10.1007/s40747-017-0064-6

Martinez, E., Calderon, C., Garcia, H., and Arguello, H. (2020). "Mri Brain Tumour Segmentation Using a Cnn over a Multi-Parametric Feature Extraction," in 2020 IEEE Colombian Conference on Applications of Computational Intelligence (IEEE ColCACI 2020), 1–6. doi:10.1109/ColCACI50549.2020.9247926

Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). "V-net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in 2016 Fourth International Conference on 3D Vision (3DV). doi:10.1109/3dv.2016.79

Minz, A., and Mahobiya, C. (2017). "Mr Image Classification Using Adaboost for Brain Tumor Type," in 2017 IEEE 7th International Advance Computing Conference (IACC), 701–705. doi:10.1109/IACC.2017.0146

[Dataset] Neelakantan, A., Vilnis, L., Le, Q. V., Sutskever, I., Kaiser, L., Kurach, K., et al. (2015). *Adding Gradient Noise Improves Learning for Very Deep Networks*.

Pembury Smith, M. Q. R., Ruxton, G. D., and Ruxton, G. D. (2020). Effective Use of the Mcnemar Test. *Behav. Ecol. Sociobiol.* 74. doi:10.1007/s00265-020-02916-y

Ranjbarzadeh, R., Bagherian Kasgari, A., Jafarzadeh Ghoushchi, S., Anari, S., Naseri, M., and Bendechache, M. (2021). Brain Tumor Segmentation Based on Deep Learning and an Attention Mechanism Using MRI Multi-Modalities Brain Images. *Sci. Rep.* 11. doi:10.1038/s41598-021-90428-8

[Dataset] Ressler, K. J., and Williams, L. M. (2020). Big Data in Psychiatry: Multiomics, Neuroimaging, Computational Modeling, and Digital Phenotyping. *Neuropsychopharmacol.* 46, 1–2. doi:10.1038/s41386-020-00862-x

Richard, A., Mayag, B., Talbot, F., Tsoukias, A., and Meinard, Y. (2020). Transparency of Classification Systems for Clinical Decision Support. *Inf. Process. Manage. Uncertainty Knowledge-Based Syst.* 1239, 99–113. doi:10.1007/978-3-030-50153-2_8

Rundo, L., Militello, C., Vitabile, S., Russo, G., Sala, E., and Gilardi, M. C. (2019). A Survey on Nature-Inspired Medical Image Analysis: A Step Further in Biomedical Data Integration. *Fi* 171, 345–365. doi:10.3233/FI-2020-1887

Shahzadi, I., Tang, T. B., Meriadeau, F., and Quyyum, A. (2018). "Cnn-lstm: Cascaded Framework for Brain Tumour Classification," in 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES), 633–637. doi:10.1109/iecbes.2018.8626704

Sharma, Y., Verma, A., Rao, K., Eluri, V., Verma, A., Rao, K., et al. (2020). *'reasonable Explainability' for Regulating Ai in Health*.

Spatharou, A., Hieronimus, S., and Jenkins, J. (2021). *Transforming Healthcare with Ai: The Impact on the Workforce and Organizations*.

Sun, Y., Zhu, L., Wang, G., and Zhao, F. (2017). Multi-input Convolutional Neural Network for Flower Grading. *J. Electr. Comput. Eng.* 2017, 1–8. doi:10.1155/2017/9240407

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Vedaldi, A., and Soatto, S. (2008). "Quick Shift and Kernel Methods for Mode Seeking," in *Computer Vision – ECCV 2008*. Editors D. Forsyth, P. Torr, and A. Zisserman (Berlin, Heidelberg: Springer Berlin Heidelberg), 705–718. doi:10.1007/978-3-540-88693-8_52

Weerakkody, Y., and Gaillard, F. (2021). *Pituitary Adenoma: Radiology Reference Article*. doi:10.53347/rID-11024

Wojciech, S., Thomas, W., and Robert, M. K. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*.

Zucco, C., Liang, H., Fatta, G. D., and Cannataro, M. (2018). "Explainable Sentiment Analysis with Applications in Medicine," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 1740–1747. doi:10.1109/BIBM.2018.8621359

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 16 March 2022
doi: 10.3389/fgene.2022.862105

Identification of Potential Diagnostic and Prognostic Biomarkers for Gastric Cancer Based on Bioinformatic Analysis

Xiaoji Niu[1,2], *Liman Ren*[3], *Aiyan Hu*[2], *Shuhui Zhang*[2]* *and Hongjun Qi*[1]*

[1]Department of Gastroenterology of Traditional Chinese Medicine, Qinghai Province Hospital of Traditional Chinese Medicine, Xining, China, [2]Department of Pathology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Department of Endocrinology, Qinghai Province Hospital of Traditional Chinese Medicine, Xining, China

Background: Gastric cancer (GC) is one of the most prevalent cancers all over the world. The molecular mechanisms of GC remain unclear and not well understood. GC cases are majorly diagnosed at the late stage, resulting in a poor prognosis. Advances in molecular biology techniques allow us to get a better understanding of precise molecular mechanisms and enable us to identify the key genes in the carcinogenesis and progression of GC.

Methods: The present study used datasets from the GEO database to screen differentially expressed genes (DEGs) between GC and normal gastric tissues. GO and KEGG enrichments were utilized to analyze the function of DEGs. The STRING database and Cytoscape software were applied to generate protein–protein network and find hub genes. The expression levels of hub genes were evaluated using data from the TCGA database. Survival analysis was conducted to evaluate the prognostic value of hub genes. The GEPIA database was involved to correlate key gene expressions with the pathological stage. Also, ROC curves were constructed to assess the diagnostic value of key genes.

Results: A total of 607 DEGs were identified using three GEO datasets. GO analysis showed that the DEGs were mainly enriched in extracellular structure and matrix organization, collagen fibril organization, extracellular matrix (ECM), and integrin binding. KEGG enrichment was mainly enriched in protein digestion and absorption, ECM-receptor interaction, and focal adhesion. Fifteen genes were identified as hub genes, one of which was excluded for no significant expression between tumor and normal tissues. COL1A1, COL5A2, P4HA3, and SPARC showed high values in prognosis and diagnosis of GC.

Conclusion: We suggest COL1A1, COL5A2, P4HA3, and SPARC as biomarkers for the diagnosis and prognosis of GC.

Keywords: gastric cancer, bioinformatics analysis, microarray, differentially expressed genes, prognosis, diagnosis

Edited by:
Saurav Mallik,
Harvard University, United States

Reviewed by:
Shima Sadri,
Marquette University, United States
Aimin Li,
Xi'an University of Technology, China

Correspondence:
Shuhui Zhang
shzhang@126.com
Hongjun Qi
docqhj@sina.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *25 January 2022*
Accepted: *25 February 2022*
Published: *16 March 2022*

Citation:
Niu X, Ren L, Hu A, Zhang S and Qi H (2022) Identification of Potential Diagnostic and Prognostic Biomarkers for Gastric Cancer Based on Bioinformatic Analysis. Front. Genet. 13:862105. doi: 10.3389/fgene.2022.862105

INTRODUCTION

According to data published in 2021, gastric cancer (GC), among all cancers, ranked fourth in cancer-related deaths (Sung et al., 2021). Stomach adenocarcinoma (STAD), the most common histological type, accounts for more than 90% of GC (Ajani et al., 2017). Although endoscopy or histological detection has developed a lot in recent years, the majority of GC patients are diagnosed at their late and advanced stage due to an insidious onset, resulting in high morbidity and mortality (Chen et al., 2020; Wang W. et al., 2020). However, advances in molecular biology techniques allow us to approach precise molecular mechanisms of carcinogenesis and enable us to find potential diagnostic and prognostic biomarkers for GC.

Previous bioinformatic studies resulted in different biomarkers due to different screening criteria and different datasets from Gene Expression Omnibus (GEO) (Sun et al., 2017; Zheng H.-C. et al., 2017; Shi and Zhang, 2019). In the present study, we identified DEGs based on three datasets from GEO, GSE19826, GSE54129, and GSE118916. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed subsequently. Afterwards, we constructed the protein–protein interaction (PPI) network to identify hub genes using the STRING database and Cytoscape software. Then, we performed the survival analysis, including overall survival (OS), disease-free survival (DSS), and progress-free interval (PFI) to identify candidate genes. The expression of candidate genes and their correlation with the pathological stage were further analyzed along with the diagnostic value. A total of four genes were identified as potential biomarkers for GC in our study.

MATERIALS AND METHODS

Microarray Datasets

RNA-sequencing datasets containing gastric cancer tissue samples and normal tissue samples were obtained from the GEO database (Barrett et al., 2013), (https://www.ncbi.nlm.nih.gov/geo/) and three GEO datasets, including GSE19826 (Wang Q. et al., 2012), GSE54129, and GSE118916 (Li et al., 2019), were downloaded for further analysis.

Identification of Differentially Expressed Genes

The limma package (version: 3.40.2) of R software was used to identity the DEGs in three datasets. The adjusted p value was

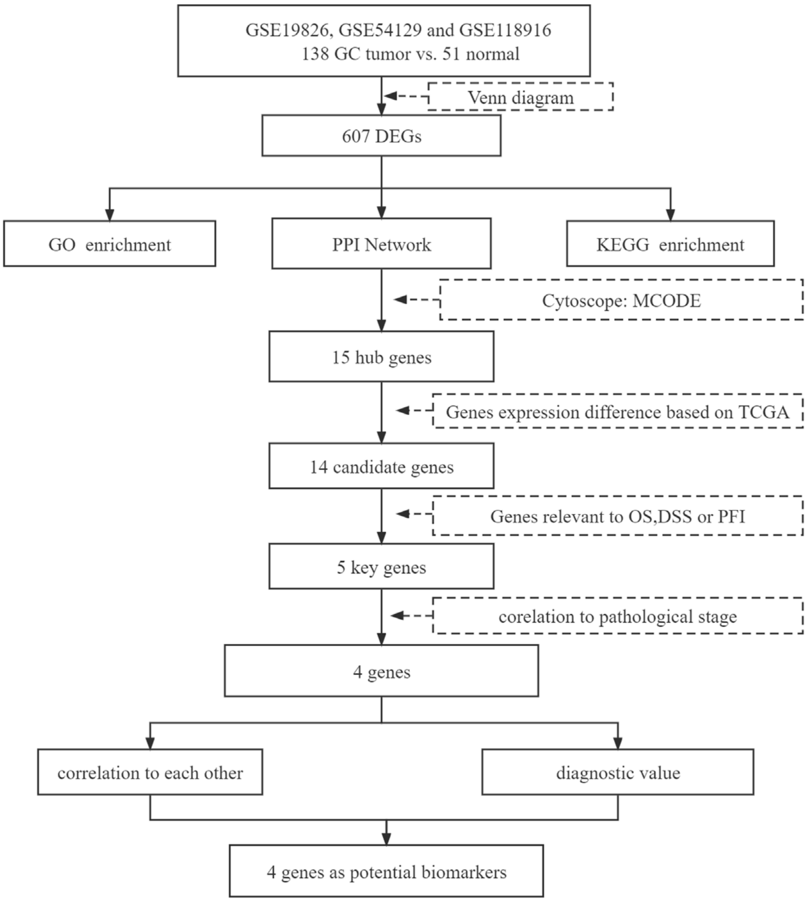

GSE19826, GSE54129 and GSE118916
138 GC tumor vs. 51 normal
Venn diagram
607 DEGs
GO enrichment
PPI Network
KEGG enrichment
Cytoscope: MCODE
15 hub genes
Genes expression difference based on TCGA
14 candidate genes
Genes relevant to OS,DSS or PFI
5 key genes
corelation to pathological stage
4 genes
correlation to each other
diagnostic value
4 genes as potential biomarkers

FIGURE 1 | Flowchart diagram for bioinformatics analysis.

analyzed to correct for false positive results in GEO datasets. "Adjusted $p < 0.05$ and fold change >1.5" were defined as the thresholds for the screening of the differential expression of mRNAs. Subsequently, the ggplot package (version: 3.3.3) of R software was used to make a Venn diagram to extract the common DEGs of the three datasets.

Enrichment Analysis of Differentially Expressed Genes

The clusterProfiler package (Yu et al., 2012) (version 3.14.3) of R software was used for enrichment analysis with the following ontology sources: GO biological processes (BPs), cellular components (CCs), molecular functions (MFs), and KEGG pathway. Adjusted $p < 0.05$ and $q < 0.2$ were set as the critical standard for significant enrichment.

Analysis of Protein–Protein Interaction Network

The PPI network of DEGs was generated using the search tool of the STRING database (Szklarczyk et al., 2019) (version 11.5). The "Multiple Proteins by Names/Identifiers" tool was chosen in this study. The organism was set as "*Homo Sapiens*." Required score was set as high confidence (0.700), and FDR stringency was set to medium (5%). The PPI network was exported for further analysis with the Cytoscape software (Otasek et al., 2019) (version 3.8.2). The plugin MCODE (Bandettini et al., 2012) (version 2.0.0, degree cutoff: 2, node score cutoff: 0.2, K-score: 2) was applied to identify the hub genes in the PPI network. The module with the highest degree was used in the following analysis.

The Expressions and Survival Analysis of Hub Genes

The Cancer Genome Atlas (TCGA) project is an open database aiming to link cancer genomic data to patients' clinicopathological information (https://www.cancer.gov/tcga). Raw counts of RNA-sequencing data (level 3) were obtained from TCGA along with corresponding clinicopathological information (Liu et al., 2018). TPM-formatted RNA-sequencing data of normal tissues from the Genotype-Tissue Expression Project (GTEx) were obtained from the University of California Santa Cruz (Vivian et al., 2017) (https://xenabrowser.net/datapages/). Tumor/normal differential expression analyses of hub genes were conducted using R software. We conducted the survival analysis, including the OS, DSS, and PFI, with the Xiantao Academic platform (survminer package of R software). DEGs related to the OS, DSS, and PFI were considered as our purpose genes and were involved in the following data analysis.

Correlation Analysis of Purpose Genes

GEPIA (Tang et al., 2017) (http://gepia.cancer-pku.cn/) is a database that enable users to analyze the RNA-sequencing expression in various ways. We used GEPIA to correlate our purpose genes with the pathological stage. Correlation analysis among purpose genes were conducted using R software embedded in Xiantao Academic. Correlation among these purpose genes were visualized with a heat map generated by the ggplot package.

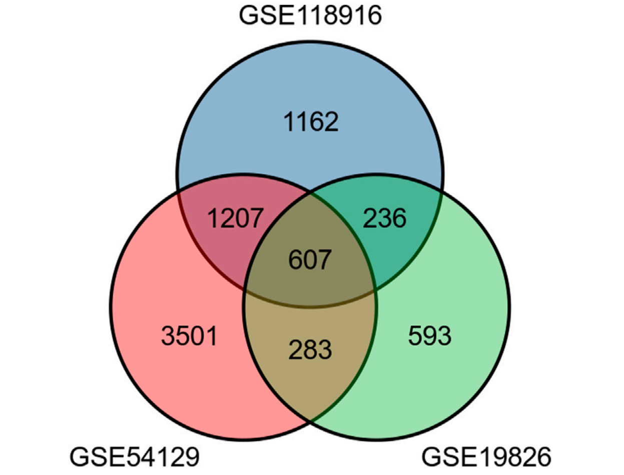


FIGURE 2 | The Venn diagram shows a total of 607 differentially expressed genes including 294 up-regulated genes and 313 down-regulated genes.

Statistical Analysis

Xiantao Academic (https://www.xiantao.love/products) is a platform embedded with R software and R packages for data analyzing. The major analysis was performed using Xiantao Academic in the present study. Chi-square test and the Wilcoxon rank sum test were utilized in the analysis depending on the data. Spearman correlation analysis was used in different expression of genes. In the analysis of the correlation of gene expression with pathological stage, the expression data are first log2 (TPM+1) transformed, and the method was one-way ANOVA, using pathological stage as a variable for calculating differential expression. p value <0.05 was regarded as statistically significant.

RESULTS

Identification of Differentially Expressed Genes

The present study involved three GEO datasets, GSE19826, GSE54129, and GSE118916. GSE19826 contained 12 pairs of samples from GC tumor and adjacent non-tumor tissues and three normal tissues. GSE54129 contained 111 GC tumor tissue samples and 21 normal tissue samples. GSE118916 contained 15 pairs of Gastric cancer tumor and adjacent non-tumor (normal) tissues. There were 138 GC tumor tissue samples and 51 normal tissue samples in total involved in the present study. The flow chart is shown in **Figure 1**. We identified 607 DEGs including 294 up-regulated genes and 313 down-regulated genes in GC tissue samples (**Figure 2**).

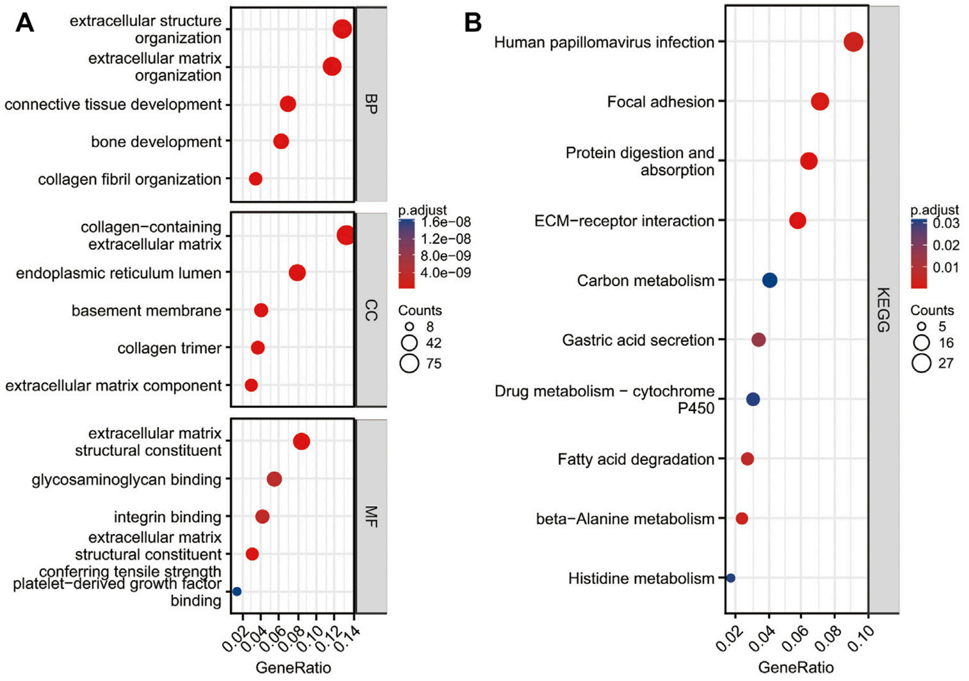


FIGURE 3 | Functional analysis of DEGs. Top five GO terms enrichment in biological process (BP), cell composition (CC), and molecular function (MF) **(A)**. KEGG enrichment of DGEs **(B)**.

Functional Enrichment Analysis of Differentially Expressed Genes

We conducted a functional enrichment analysis of DEGs using R software and R codes embedded in Xiantao platform. DEGs are enriched in 341 terms of GO BP, including extracellular structure organization, extracellular matrix (ECM) organization, collagen fibril organization, bone development and connective tissue development, etc. DEGs were enriched in 43 terms of GO CC, including collagen-containing extracellular matrix, endoplasmic reticulum lumen, basement membrane, extracellular matrix component and collagen trimer, etc. DEGs were enriched in 35 terms of GO MF, including extracellular matrix structural constituent, extracellular matrix structural constituent conferring tensile strength, integrin binding, glycosaminoglycan binding, and platelet-derived growth factor binding (**Figure 3A**; **Supplementary Table S1**). DEGs were enriched in 10 terms of KEGG, including protein digestion and absorption, ECM-receptor interaction, Focal adhesion, human papillomavirus infection, beta-alanine metabolism, fatty acid degradation, gastric acid secretion, histidine metabolism, drug metabolism—cytochrome P450, and carbon metabolism (**Figure 3B**; **Supplementary Table S1**).

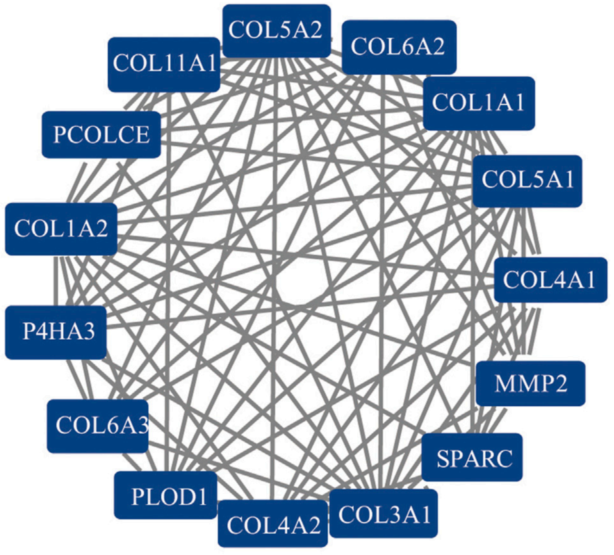


FIGURE 4 | The PPI network of 15 hub genes selected with the MCODE plugin of Cytoscape.

Protein–Protein Interaction Network to Identify Hub Genes

A PPI network of 607 DEGs, containing a total of 317 nodes and 606 edges, was generated using STRING, and an interaction score >0.7 was considered a high-confidence interaction relationship. We identified 15 nodes and 81 edges with MCODE plugin. The module with the highest degree was used in the following analysis. The hub genes included COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A2, COL6A3, COL11A1,

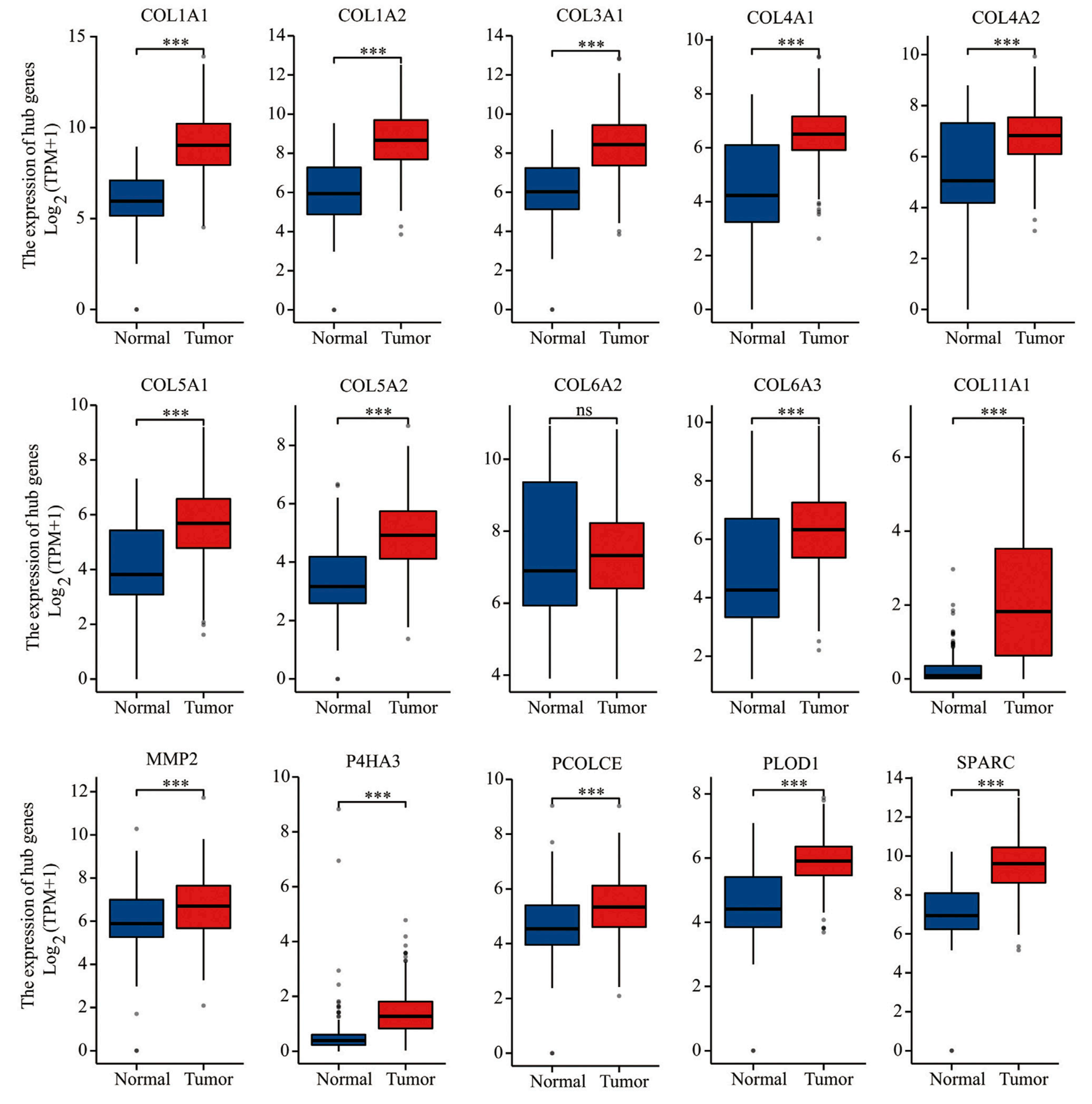


FIGURE 5 | Gene expression of 15 hub genes (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A2, COL6A3, COL11A1, MMP2, P4HA3, PCOLCE, PLOD1, and SPARC) based on TGCA and GTEx databases. ***p < 0.001; ns, not statistically significant.

MMP2, P4HA3, PCOLCE, PLOD1, and SPARC (**Figure 4**). Gene expression profiles of the 15 hub genes between GC tumor samples and normal samples are shown in **Figure 5**. The expression of COL6A2 showed no difference in tumor and normal tissues, so it was excluded in further analyses. The remaining 14 genes were considered as candidate genes for potential diagnostic and prognostic biomarkers.

Survival and Correlation Analysis

We conducted Kaplan–Meier survival analysis with the candidate genes. Candidate genes related to OS, DSS, or PFI were considered as key genes. Among the 14 candidate genes, COL1A1 (HR = 1.41, p = 0.042), COL4A1 (HR = 1.45, p = 0.029), COL5A2 (HR = 1.54, p = 0.011), P4HA3 (HR = 1.57, p = 0.011), and SPARC (HR = 1.47, p = 0.022) were associated with the OS of STAD (**Figure 6**). COL5A2 was

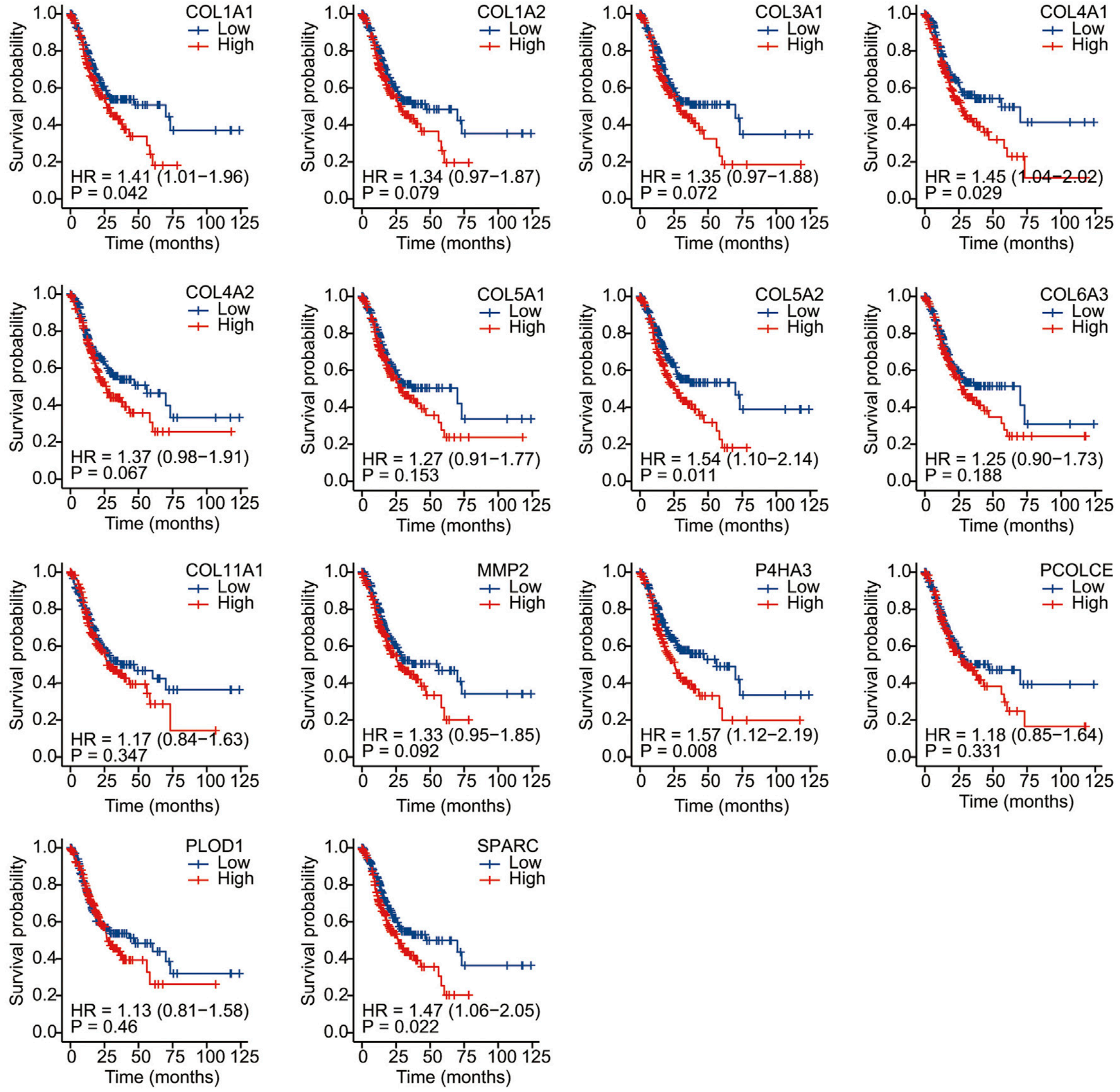


FIGURE 6 | Overall survival analysis of 14 candidate genes (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A3, COL11A1, MMP2, P4HA3, PCOLCE, PLOD1, and SPARC). COL1A1, COL4A1, COL5A2, P4HA3, and SPARC (HR = 1.47, *p* = 0.022) were associated with OS.

associated with DSS (HR = 1.70, *p* = 0.015) (**Figure 7**) and PFI (HR = 1.44, *p* = 0.043) (**Figure 8**). Therefore, this study focused on the five key genes, COL1A1, COL4A1, COL5A2, P4HA3, and SPARC. Further analysis of the correlation between these key genes and the pathological stage of GC showed that COL1A1, COL5A2, P4HA3, and SPARC were significantly correlated to cancer pathological stages. However, COL4A1 showed no significance in the correlation analysis (**Figure 9**). Therefore, we identify COL1A1, COL5A2, P4HA3, and SPARC as potential biomarkers for prognosis of GC.

Correlation Expression and Diagnostic Analysis

We analyzed the correlation between these four genes on Xiantao Academic based on data from TCGA and found that all of these genes were highly correlated with each other. The r value ranged from 0.84 to 0.92 ($p < 0.01$) (**Figure 10**). We used a receiver operating characteristic (ROC) curve to assess the diagnostic value of the purpose genes using Xiantao Academic tools based on TCGA and GTEx samples. The area under curve (AUC) of COL1A1, COL5A2, P4HA3, and SPARC was 0.916, 0.802, 0.874, and 0.895, respectively. The results, as shown previously, suggested that these four genes we selected could effectively distinguish GC samples with normal samples (**Figure 11**). COL1A1, COL5A2, P4HA3, and SPARC could be biomarkers for the diagnosis and prognosis of GC.

DISCUSSION

GC is one of the most diagnosed cancers and has brought great burden to global health. Patients were likely to be diagnosed in their late stage due to the lack of specific clinical symptoms at an early

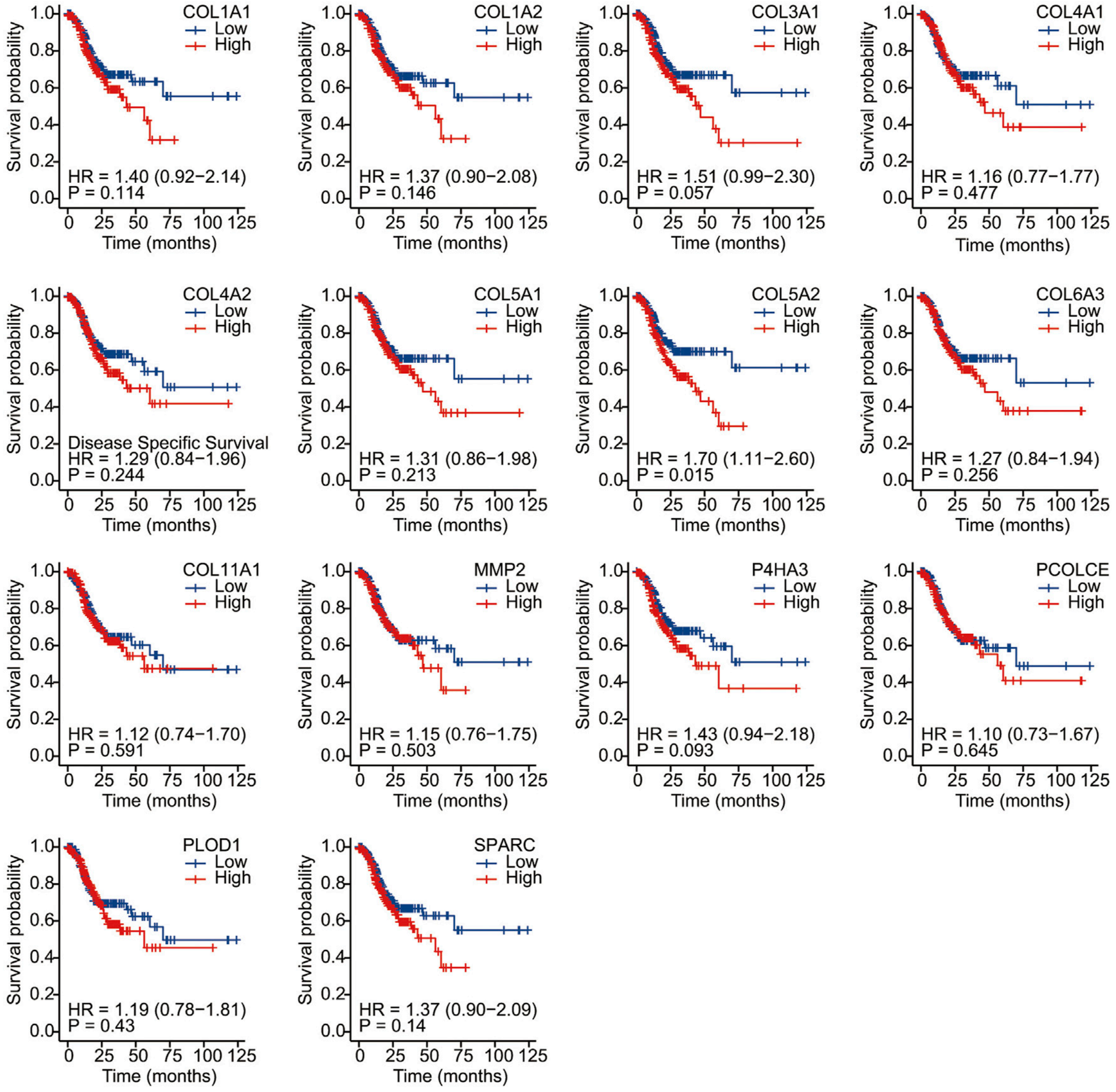


FIGURE 7 | Disease-specific survival analysis of 14 candidate genes (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A3, COL11A1, MMP2, P4HA3, PCOLCE, PLOD1, and SPARC). COL5A2 was associated with DSS (HR = 1.70, *p* = 0.015).

stage. Thus, patients with GC have poor prognosis. It is urgent to identify relevant biomarkers that are valid for both diagnostic and prognostic evaluation. Bioinformatics analysis enables us to explore the genetic alterations in GC and has been proved to be a useful approach to identify new biomarkers in plenty of diseases. An initial objective of the project was to identify appropriate biomarkers of GC using bioinformatics analysis.

In the current study, we identified 607 DEGs meeting the criteria. GO enrichment suggested those genes were significantly associated with extracellular structure and matrix organization, collagen fibril organization, and ECM and integrin binding. KEGG was mainly enriched in protein digestion and absorption, ECM-receptor interaction, and focal adhesion. In accordance with the present results, previous studies have reported that cancer-associated fibroblasts are essential in creating extracellular matrix structure and metabolism and account for the adaptive resistance to chemotherapy caused by immune reprogramming of the tumor microenvironment (Quante et al., 2011; Kalluri, 2016). Extracellular matrix plays a significant part in the creation of tumor microenvironment and promotes malignancy (Madsen and Sidenius, 2008; Najafi et al., 2019; Mohan et al., 2020; Piersma et al., 2020; Wang W. et al., 2020). Integrins coordinate ECM–cell and cell–cell interactions, signal transmission, gene expression, and cell function. The interaction between integrin and the cancer glycol microenvironment plays a significant part in regulating cancer progression (Marsico et al., 2018).

The results of this study showed that COL1A1, COL4A1, COL5A2, P4HA3, and SPARC were associated with the OS of GC. COL5A2 was associated with DSS and PFI. Further analysis of the correlation between these key genes and the pathological

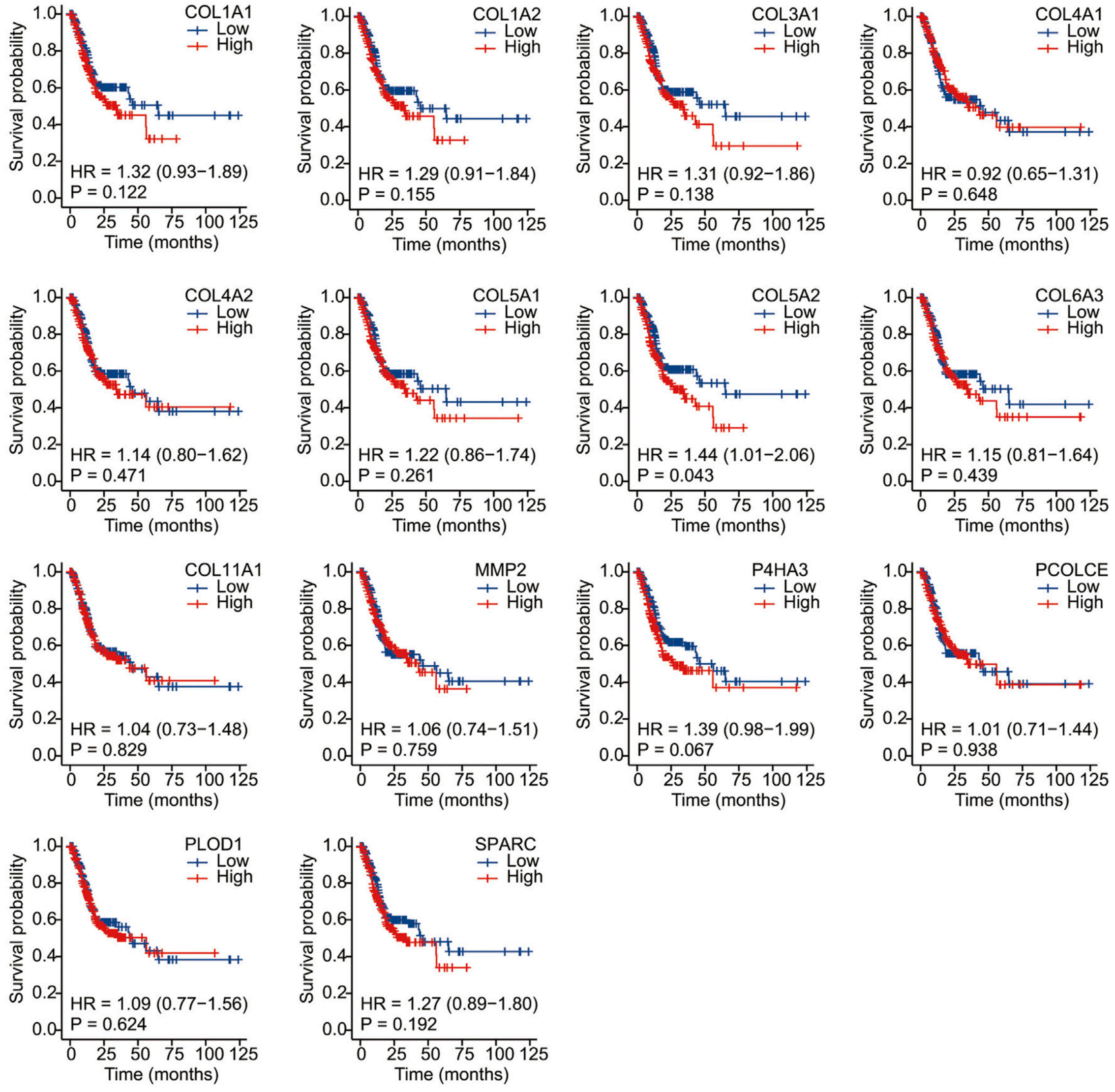


FIGURE 8 | Progress free interval analysis of 14 candidate genes (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL5A1, COL5A2, COL6A3, COL11A1, MMP2, P4HA3, PCOLCE, PLOD1, and SPARC). COL5A2 was associated with PFI (HR = 1.44, *p* = 0.043).

stages of GC showed that COL4A1 showed no significance in the correlation analysis to pathological stages. Therefore, we identify four genes as potential biomarkers of GC including COL1A1, COL5A2, P4HA3, and SPARC. Also, the diagnostic value of these genes was confirmed in the following analysis.

COL1A1 is an important member of the type-I collagen family, the main fibrillar collagen and an essential structural component of the ECM (Li J. et al., 2016). Many bioinformatic analyses identified COL1A1 as a biomarker of GC (Wang W. et al., 2020; Wang Y. et al., 2021; Zhao et al., 2021). Abnormal expression of COL1A1 has been reported in several cancers, including hepatocellular carcinoma, ovarian cancer, and colorectal cancer, as well as in GC (Li J. et al., 2016; Zhang et al., 2018; Ma et al., 2019; Li et al., 2020). *In vitro*, enhanced expression of COL1A1 promotes the invasion and migration of GC cells, while knocking out COL1A1 inhibits the increase in cell metastasis ability (Li et al., 2021). It plays an important role in promoting tumor cell proliferation, migration, invasion, epithelial–mesenchymal transformation (EMT), and chemotherapy resistance (Armstrong et al., 2004; Koenig et al., 2006; Shintani et al., 2008; Yang et al., 2014; Zheng X. et al., 2017; Yamazaki et al., 2018; Shi et al., 2021). ROC analysis showed high diagnostic value of COL1A1 (AUC = 0.916) based on 414 GC samples and 210 normal gastric tissues. This finding is consistent with that of Zhao et al. (2021) (AUC = 0.917) based on 375 GC samples and 32 normal samples (Zhao et al., 2021). The diagnostic and prognostic values of COL1A1 were confirmed with extra data (more samples than others) from the present work.

COL5A2 is a member of the type-V collagen family which is also a significant structural component of the ECM. COL5A2 was

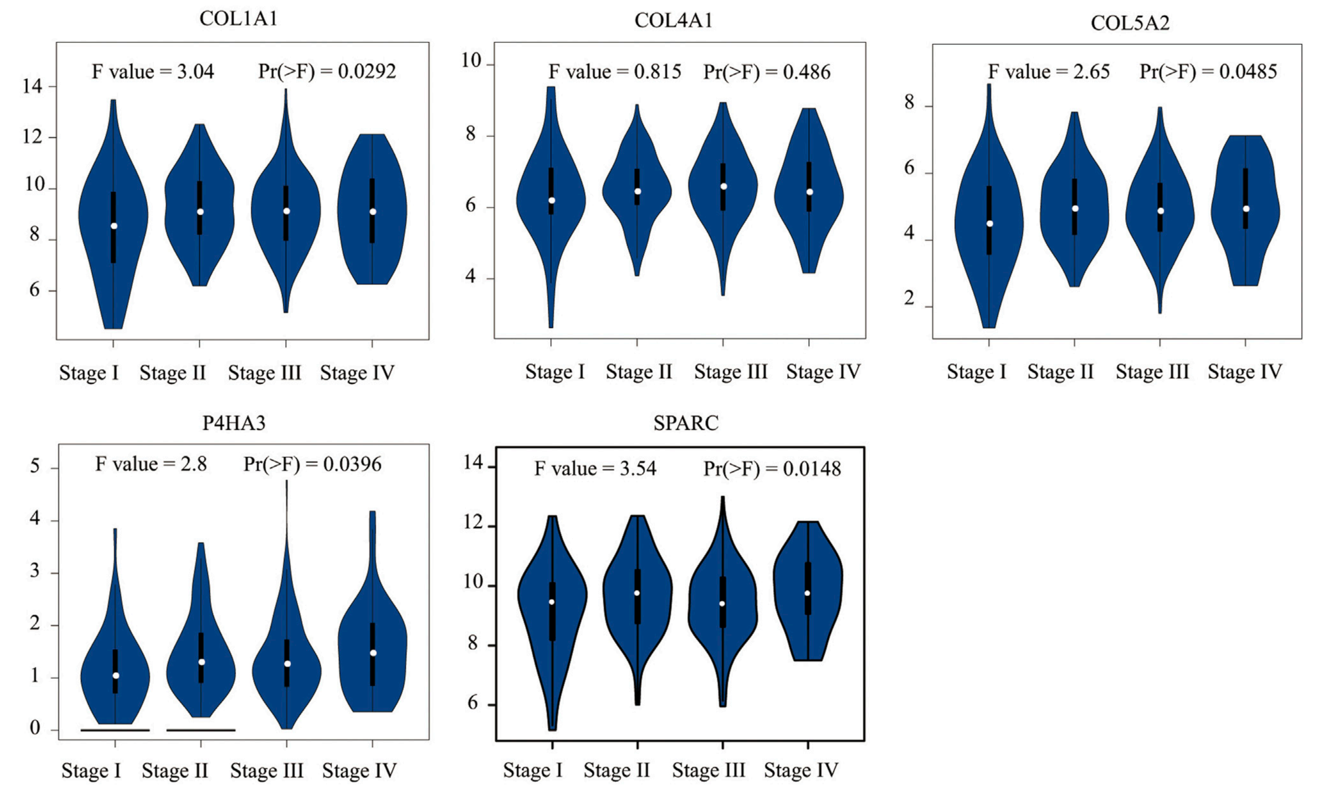


FIGURE 9 | Correlation analysis between five key genes (COL1A1, COL4A1, COL5A2, P4HA3, and SPARC) and the pathological stage of GC shows they are potential prognostic markers.

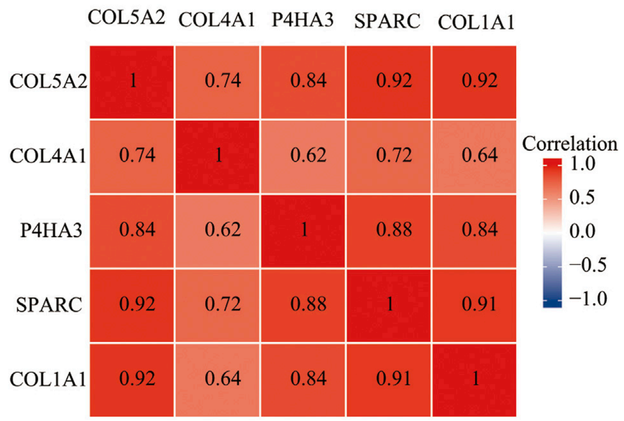


FIGURE 10 | The expression of four genes (COL1A1, COL5A2, P4HA3, and SPARC) are correlated with each other in GC.

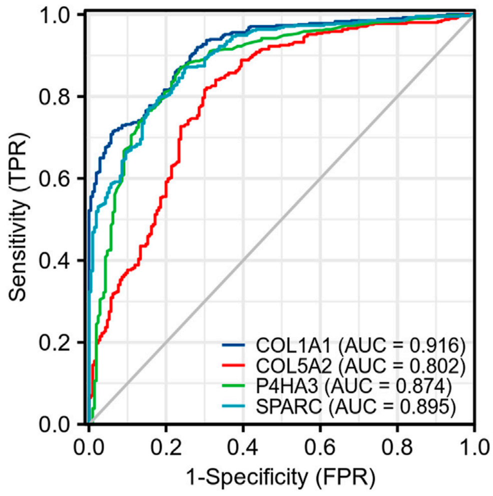


FIGURE 11 | ROC of four key genes (COL1A1, COL5A2, P4HA3, and SPARC) shows they are of high diagnostic value in GC.

reported to promote proliferation and invasion in colon cancer and prostate cancer (Ren et al., 2021; Wang J. et al., 2021). Also, it has a strong correlation to the prognosis of renal cancer and gastric cancer (Ding et al., 2021; Tan et al., 2021). The overexpression of COL5A2 promoted the migration of GC cells *in vitro* and *in vivo*, and the knockdown of COL5A2 could significantly decrease the migration of cell (Tan et al., 2021). A previous study had demonstrated that patients with higher COL5A2 levels were more likely to suffer from renal metastasis (AUC = 0.878). Among all those genes we identified as potential biomarkers, COL5A2 was the unique gene that was associated with the OS, DSS, and PFI of GC, which had not been reported in previous studies. The value of AUC in our current study is 0.802 based on data from TCGA and GTEx. Therefore, COL5A2 could serve as a novel biomarker of GC. Also, we would perform biological experiments to verify the result.

Previous research showed that P4HA3 was up-regulated in head and neck squamous cell carcinoma (HNSCC) tissue, and it was demonstrated to promote HNSCC cell proliferation, invasion, and migration *in vitro* (Wang T. et al., 2020). A recent study showed that the de-regulation of P4HA3 was associated with increased metastasis and poor prognosis of GC (Song et al., 2018). In the present work, the value of AUC of P4HA3 is 0.875, which indicated high value of diagnosis and has not been reported in previous studies. The result suggests that P4HA3 is a potential biomarker of GC.

SPARC is one of the first-known matricellular protein that modulates interactions between cells and the ECM. It has divergent actions due to different categories of tumors. It shows anti-tumor or tumor-promoting effects in different cancers (Tai and Tang, 2008). What is surprising is that previous research results are inconsistent. As Zhang et al. (2012) and Zhang et al. (2014) reported, "SPARC expression is negatively correlated with the clinicopathological factors of gastric cancer and inhibits malignancy of gastric cancer cells," and they confirmed the anti-tumor activity of SPARC *in vivo* and *in vitro*. The anti-tumor activity was also reported by Wang L. et al. (2012) in a clinical trial involving 80 gastric cancer samples and 30 normal samples. On the contrary, the tumor-promoting effect of SPARC was also reported in GC. Over expression of SPARC promoted GC progression, including serosal invasion, lymph node, and distant metastasis, and tended to poor prognosis of patients (Zhao et al., 2010; Sato et al., 2013; Wang et al., 2014). Also, the invasion and proliferation ability was inhibited in SPARC knockdown MGC803 and HGC 27 gastric cancer cell lines, which demonstrated the tumor-promoting activity of SPARC. Increased expression of SPARC in this study corroborates these earlier findings (Li Z. et al., 2016; Liao et al., 2018; Li et al., 2019). ROC analysis showed high diagnostic value of SPARC, and the value of AUC was 0.895 in the current study. Biological experiments in different cell lines and clinical samples are necessary to verify the result.

The correlation between these four genes was analyzed, and we found that all of these genes were highly correlated with each other, which enhanced their possibility as potential biomarkers of GC.

In summary, previous studies have identified COL1A1 as a biomarker for GC diagnosis and prognosis. COL5A2, P4HA3, and SPARC were reported to be associated with poor prognosis (OS and DSS, but not PFI); however, the diagnostic value has not been recognized. In the present study, the prognosis values of COL1A1, COL5A2, P4HA3, and SPARC were confirmed. The ROC analysis showed that they could distinguish between GC samples and normal samples effectively. Thus, we suggest COL1A1, COL5A2, P4HA3, and SPARC as biomarkers for both diagnosis and prognosis of GC. Each of the biomarkers identified in the present work plays a significant role in the ECM, which highlights the importance of the tumor microenvironment in GC. Compared with similar studies, we suggested those genes as both diagnostic and prognostic biomarkers for GC. Nevertheless, the current results are all derived from bioinformatics analysis and are limited by the absence of confirmation. Due to different screening criteria, previous bioinformatics research produced different biomarkers. Many of the biomarkers have been verified, and the combination of those results might be more rigorous. Further clinical experiments are underway to verify their value in GC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. These data can be found at https://www.ncbi.nlm.nih.gov/geo/download/?acc = GSE19826; https://www.ncbi.nlm.nih.gov/geo/download/?acc = GSE54129; and https://www.ncbi.nlm.nih.gov/geo/download/?acc = GSE118916.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Qinghai Province Hospital of Traditional Chinese Medicine. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

XN, SZ, and HQ contributed to conception and design of the study. XN and LR organized the database and performed the statistical analysis. XN wrote the first draft of the manuscript. AH wrote sections of the manuscript. SZ and HQ contributed to manuscript revision. All authors read and approved the submitted version.

FUNDING

The present work was funded by the Construction Project of Provincial Clinical Key Subject from the Health Commission of Qinghai Province.

ACKNOWLEDGMENTS

The authors would like to thank TCGA, GEO databases, and the University of California Santa Cruz for providing platforms and contributors for uploading their meaningful datasets. They thank Xiantao Academic for their contribution in data analyzing and data visualizing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.862105/full#supplementary-material

REFERENCES

Ajani, J. A., Lee, J., Sano, T., Janjigian, Y. Y., Fan, D., and Song, S. (2017). Gastric Adenocarcinoma. *Nat. Rev. Dis. Primers* 3 (1), 17036. doi:10.1038/nrdp.2017.36

Armstrong, T., Packham, G., Murphy, L. B., Bateman, A. C., Conti, J. A., Fine, D. R., et al. (2004). Type I Collagen Promotes the Malignant Phenotype of Pancreatic Ductal Adenocarcinoma. *Clin. Cancer Res.* 10 (21), 7427–7437. doi:10.1158/1078-0432.CCR-03-0825

Bandettini, W. P., Kellman, P., Mancini, C., Booker, O. J., Vasu, S., Leung, S. W., et al. (2012). MultiContrast Delayed Enhancement (MCODE) Improves Detection of Subendocardial Myocardial Infarction by Late Gadolinium Enhancement Cardiovascular Magnetic Resonance: A Clinical Validation Study. *J. Cardiovasc. Magn. Reson.* 14 (1), 83. doi:10.1186/1532-429X-14-83

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: Archive for Functional Genomics Data Sets-Update. *Nucleic Acids Res.* 41 (D1), D991–D995. doi:10.1093/nar/gks1193

Chen, Y., Chen, W., Dai, X., Zhang, C., Zhang, Q., and Lu, J. (2020). Identification of the Collagen Family as Prognostic Biomarkers and Immune-Associated Targets in Gastric Cancer. *Int. Immunopharmacol.* 87, 106798. doi:10.1016/j.intimp.2020.106798

Ding, Y.-L., Sun, S.-F., and Zhao, G.-L. (2021). COL5A2 as a Potential Clinical Biomarker for Gastric Cancer and Renal Metastasis. *Medicine* 100 (7), e24561. doi:10.1097/MD.0000000000024561

Kalluri, R. (2016). The Biology and Function of Fibroblasts in Cancer. *Nat. Rev. Cancer* 16, 582–598. doi:10.1038/nrc.2016.73

Koenig, A., Mueller, C., Hasel, C., Adler, G., and Menke, A. (2006). Collagen Type I Induces Disruption of E-Cadherin-Mediated Cell-Cell Contacts and Promotes Proliferation of Pancreatic Carcinoma Cells. *Cancer Res.* 66 (9), 4662–4671. doi:10.1158/0008-5472.CAN-05-2804

Li, J., Ding, Y., and Li, A. (2016). Identification of COL1A1 and COL1A2 as Candidate Prognostic Factors in Gastric Cancer. *World J. Surg. Onc* 14 (1), 297. doi:10.1186/s12957-016-1056-5

Li, L., Zhu, Z., Zhao, Y., Zhang, Q., Wu, X., Miao, B., et al. (2019). FN1, SPARC, and SERPINE1 are Highly Expressed and Significantly Related to a Poor Prognosis of Gastric Adenocarcinoma Revealed by Microarray and Bioinformatics. *Sci. Rep.* 9 (1), 7827. doi:10.1038/s41598-019-43924-x

Li, M., Wang, J., Wang, C., Xia, L., Xu, J., Xie, X., et al. (2020). Microenvironment Remodeled by Tumor and Stromal Cells Elevates Fibroblast-Derived COL1A1 and Facilitates Ovarian Cancer Metastasis. *Exp. Cel Res.* 394 (1), 112153. doi:10.1016/j.yexcr.2020.112153

Li, Y., Sun, R., Zhao, X., and Sun, B. (2021). RUNX2 Promotes Malignant Progression in Gastric Cancer by Regulating COL1A1. *Cancer Biomarkers* 31 (3), 1–12. doi:10.3233/CBM-200472

Li, Z., Z., LiLi, A.-D., Xu, L., Bai, D.-W., Hou, K.-Z., Zheng, H.-C., et al. (2016). SPARC Expression in Gastric Cancer Predicts Poor Prognosis: Results from a Clinical Cohort, Pooled Analysis and GSEA Assay. *Oncotarget* 7 (43), 70211–70222. doi:10.18632/oncotarget.12191

Liao, P., Li, W., Liu, R., Teer, J. K., Xu, B., Zhang, W., et al. (2018). Genome-Scale Analysis Identifies SERPINE1 and SPARC as Diagnostic and Prognostic Biomarkers in Gastric Cancer. *OncoTargets Ther.* 11, 6969–6980. doi:10.2147/OTT.S173934

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 173 (2), 400–e11. doi:10.1016/j.cell.2018.02.052

Ma, H.-P., Chang, H.-L., Bamodu, O. A., Yadav, V. K., Huang, T.-Y., Wu, A. T. H., et al. (2019). Collagen 1A1 (COL1A1) Is a Reliable Biomarker and Putative Therapeutic Target for Hepatocellular Carcinogenesis and Metastasis. *Cancers* 11 (6), 786. doi:10.3390/cancers11060786

Madsen, C. D., and Sidenius., N. (2008). The Interaction between Urokinase Receptor and Vitronectin in Cell Adhesion and Signalling. *Eur. J. Cell Biol.* 87, 617–629. doi:10.1016/j.ejcb.2008.02.003

Marsico, G., Russo, L., Quondamatteo, F., and Pandit, A. (2018). Glycosylation and Integrin Regulation in Cancer. *Trends Cancer* 4, 537–552. doi:10.1016/j.trecan.2018.05.009

Mohan, V., Das, A., and Sagi, I. (2020). Emerging Roles of ECM Remodeling Processes in Cancer. *Semin. Cancer Biol.* 62, 192–200. doi:10.1016/j.semcancer.2019.09.004

Najafi, M., Farhood, B., and Mortezaee, K. (2019). Extracellular Matrix (ECM) Stiffness and Degradation as Cancer Drivers. *J. Cel Biochem* 120 (3), 2782–2790. doi:10.1002/jcb.27681

Otasek, D., Morris, J. H., Bouças, J., Pico, A. R., and Demchak, B. (2019). Cytoscape Automation: Empowering Workflow-Based Network Analysis. *Genome Biol.* 20 (1), 185. doi:10.1186/s13059-019-1758-4

Piersma, B., Hayward, M. K., and Weaver, V. M. (2020). Fibrosis and Cancer: A Strained Relationship. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1873, 188356. doi:10.1016/j.bbcan.2020.188356

Quante, M., Tu, S. P., Tomita, H., Gonda, T., Wang, S. S. W., Takashi, S., et al. (2011). Bone Marrow-Derived Myofibroblasts Contribute to the Mesenchymal Stem Cell Niche and Promote Tumor Growth. *Cancer Cell* 19 (2), 257–272. doi:10.1016/j.ccr.2011.01.020

Ren, X., Chen, X., Fang, K., Zhang, X., Wei, X., Zhang, T., et al. (2021). COL5A2 Promotes Proliferation and Invasion in Prostate Cancer and is One of Seven Gleason-Related Genes that Predict Recurrence-Free Survival. *Front. Oncol.* 11, 583083. doi:10.3389/fonc.2021.583083

Sato, T., Oshima, T., Yamamoto, N., Yamada, T., Hasegawa, S., Yukawa, N., et al. (2013). Clinical Significance of SPARC Gene Expression in Patients with Gastric Cancer. *J. Surg. Oncol.* 108 (6), 364–368. doi:10.1002/jso.23425

Shi, R., Gao, S., Zhang, J., Xu, J., Graham, L. M., Yang, X., et al. (2021). Collagen Prolyl 4-Hydroxylases Modify Tumor Progression. *Acta Biochim. Biophys. Sinica* 53, 805–814. doi:10.1093/abbs/gmab065

Shi, S., and Zhang, Z. G. (2019). Role of Sp1 Expression in Gastric Cancer: A Meta-Analysis and Bioinformatics Analysis. *Oncol. Lett.* 18 (4), 4126–4135. doi:10.3892/ol.2019.10775

Shintani, Y., Maeda, M., Chaika, N., Johnson, K. R., and Wheelock, M. J. (2008). Collagen I Promotes Epithelial-to-Mesenchymal Transition in Lung Cancer Cells via Transforming Growth Factor-β Signaling. *Am. J. Respir. Cel Mol Biol* 38 (1), 95–104. doi:10.1165/rcmb.2007-0071OC

Song, H., Liu, L., Song, Z., Ren, Y., Li, C., and Huo, J. (2018). P4HA3is Epigenetically Activated by Slug in Gastric Cancer and its Deregulation is Associated with Enhanced Metastasis and Poor Survival. *Technol. Cancer Res. Treat.* 17, 153303381879648. doi:10.1177/1533033818796485

Sun, C., Yuan, Q., Wu, D., Meng, X., and Wang, B. (2017). Identification of Core Genes and Outcome in Gastric Cancer Using Bioinformatics Analysis. *Oncotarget* 8 (41), 70271–70280. doi:10.18632/oncotarget.20082

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* 47 (D1), D607–D613. doi:10.1093/nar/gky1131

Tai, I. T., and Tang, M. J. (2008). SPARC in Cancer Biology: Its Role in Cancer Progression and Potential for Therapy. *Drug Resist. Updates* 11 (6), 231–246. doi:10.1016/j.drup.2008.08.005

Tan, Y., Chen, Q., Xing, Y., Zhang, C., Pan, S., An, W., et al. (2021). High Expression of COL5A2, a Member of COL5 Family, Indicates the Poor Survival and Facilitates Cell Migration in Gastric Cancer. *Biosci. Rep.* 41 (4), BSR20204293. doi:10.1042/BSR20204293

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: A Web Server for Cancer and Normal Gene Expression Profiling and Interactive Analyses. *Nucleic Acids Res.* 45 (W1), W98–W102. doi:10.1093/nar/gkx247

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., et al. (2017). Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses. *Nat. Biotechnol.* 35, 314–316. doi:10.1038/nbt.3772

Wang, J., Jiang, Y.-H., YangYang, P.-Y., and Liu, F. (2021). Increased Collagen Type V α2 (COL5A2) in Colorectal Cancer Is Associated with Poor Prognosis and Tumor Progression. *OncoTargets Ther.* 14, 2991–3002. doi:10.2147/OTT.S288422

Wang, L., L., Yang, M., Shan, L., Qi, L., Chai, C., Zhou, Q., et al. (2012). The Role of SPARC Protein Expression in the Progress of Gastric Cancer. *Pathol. Oncol. Res.* 18 (3), 697–702. doi:10.1007/s12253-012-9497-9

Wang, Q., Wen, Y.-G., Li, D.-P., Xia, J., Zhou, C.-Z., Yan, D.-W., et al. (2012). Upregulated INHBA Expression is Associated with Poor Survival in Gastric Cancer. *Med. Oncol.* 29 (1), 77–83. doi:10.1007/s12032-010-9766-y

Wang, T., Wang, Y.-X., Dong, Y.-Q., Yu, Y.-L., and Ma, K. (2020). Prolyl 4-Hydroxylase Subunit Alpha 3 Presents a Cancer Promotive Function in Head and Neck Squamous Cell Carcinoma via Regulating Epithelial-Mesenchymal Transition. *Arch. Oral Biol.* 113, 104711. doi:10.1016/j.archoralbio.2020.104711

Wang, W., He, Y., Zhao, Q., Zhao, X., and Li, Z. (2020). Identification of Potential Key Genes in Gastric Cancer Using Bioinformatics Analysis. *Biom Rep.* 12 (4), 178–192. doi:10.3892/br.2020.1281

Wang, Y., Zheng, K., Chen, X., Chen, R., and Zou, Y. (2021). Bioinformatics Analysis Identifies COL1A1, THBS2 and SPP1 as Potential Predictors of Patient Prognosis and Immunotherapy Response in Gastric Cancer. *Biosci. Rep.* 41 (1), BSR20202564. doi:10.1042/BSR20202564

Wang, Z., Hao, B., Yang, Y., Wang, R., Li, Y., and Wu, Q. (2014). Prognostic Role of SPARC Expression in Gastric Cancer: A Meta-Analysis. *Arch. Med. Sci.* 10 (5), 863–869. doi:10.5114/aoms.2014.46207

Yamazaki, S., Higuchi, Y., Ishibashi, M., Hashimoto, H., Yasunaga, M., Matsumura, Y., et al. (2018). Collagen Type I Induces EGFR-TKI Resistance in EGFR-Mutated Cancer Cells by MTOR Activation through Akt-Independent Pathway. *Cancer Sci.* 109 (6), 2063–2073. doi:10.1111/cas.13624

Yang, M. C., Wang, C. J., Liao, P. C., Yen, C. J., and Shan, Y. S. (2014). Hepatic Stellate Cells Secretes Type I Collagen to Trigger Epithelial Mesenchymal Transition of Hepatoma Cells. *Am. J. Cancer Res.* 4 (6), 751–763.

Yu, G., Wang, L.-G., Han, Y., and HeHe, Q.-Y. (2012). ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A J. Integr. Biol.* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, J.-L., ChenChen, G.-W., LiuLiu, Y.-C., WangWang, P.-Y., Wang, X., WanWan, Y.-L., et al. (2012). Secreted Protein Acidic and Rich in Cysteine (SPARC) Suppresses Angiogenesis by Down-Regulating the Expression of VEGF and MMP-7 in Gastric Cancer. *PLoS ONE* 7 (9), e44618. doi:10.1371/journal.pone.0044618

Zhang, J., Wang, P., Zhu, J., Wang, W., Yin, J., Zhang, C., et al. (2014). SPARC Expression is Negatively Correlated with Clinicopathological Factors of Gastric Cancer and Inhibits Malignancy of Gastric Cancer Cells. *Oncol. Rep.* 31 (5), 2312–2320. doi:10.3892/or.2014.3118

Zhang, Z., Wang, Y., Zhang, J., Zhong, J., and Yang, R. (2018). COL1A1 Promotes Metastasis in Colorectal Cancer by Regulating the WNT/PCP Pathway. *Mol. Med. Rep.* 17 (4), 5037–5042. doi:10.3892/mmr.2018.8533

Zhao, Q., Xie, J., Xie, J., Zhao, R., Song, C., Wang, H., et al. (2021). Weighted Correlation Network Analysis Identifies FN1, COL1A1 and SERPINE1 Associated with the Progression and Prognosis of Gastric Cancer. *Cancer Biomarkers* 31 (1), 59–75. doi:10.3233/CBM-200594

Zhao, Z.-S., WangWang, Y.-Y., Chu, Y.-Q., Ye, Z.-Y., and TaoTao, H.-Q. (2010). SPARC is Associated with Gastric Cancer Progression and Poor Survival of Patients. *Clin. Cancer Res.* 16 (1), 260–268. doi:10.1158/1078-0432.CCR-09-1247

Zheng, H.-C., Gong, B.-C., and Zhao, S. (2017). The Meta and Bioinformatics Analysis of GRP78 Expression in Gastric Cancer. *Oncotarget* 8 (42), 73017–73028. doi:10.18632/oncotarget.20318

Zheng, X., Liu, W., Xiang, J., Liu, P., Ke, M., Wang, B., et al. (2017). Collagen I Promotes Hepatocellular Carcinoma Cell Proliferation by Regulating Integrin β1/FAK Signaling Pathway in Nonalcoholic Fatty Liver. *Oncotarget* 8 (56), 95586–95595. doi:10.18632/oncotarget.21525

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 26 April 2022
doi: 10.3389/fgene.2022.844391

An Ensemble-Based Deep Convolutional Neural Network for Computer-Aided Polyps Identification From Colonoscopy

Pallabi Sharma[1], *Bunil Kumar Balabantaray*[1], *Kangkana Bora*[2], *Saurav Mallik*[3], *Kunio Kasugai*[4] and *Zhongming Zhao*[3,5,6]*

[1]Department of Computer Science and Engineering, National Institute of Technology Meghalaya, Shillong, India, [2]Computer Science and Information Technology, Cotton University, Guwahati, India, [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States, [4]Department of Gastroenterology, Aichi Medical University, Nagakute, Japan, [5]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States, [6]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX, United States

OPEN ACCESS

Edited by:
Ka-Chun Wong,
City University of Hong Kong, Hong Kong SAR, China

Reviewed by:
Wei Zhang,
University of Central Florida, United States
Szilvia Nagy,
Széchenyi István University, Hungary

***Correspondence:**
Zhongming Zhao
zhongming.zhao@uth.tmc.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *28 December 2021*
Accepted: *14 March 2022*
Published: *26 April 2022*

Citation:
Sharma P, Balabantaray BK, Bora K, Mallik S, Kasugai K and Zhao Z (2022) An Ensemble-Based Deep Convolutional Neural Network for Computer-Aided Polyps Identification From Colonoscopy. Front. Genet. 13:844391. doi: 10.3389/fgene.2022.844391

Colorectal cancer (CRC) is the third leading cause of cancer death globally. Early detection and removal of precancerous polyps can significantly reduce the chance of CRC patient death. Currently, the polyp detection rate mainly depends on the skill and expertise of gastroenterologists. Over time, unidentified polyps can develop into cancer. Machine learning has recently emerged as a powerful method in assisting clinical diagnosis. Several classification models have been proposed to identify polyps, but their performance has not been comparable to an expert endoscopist yet. Here, we propose a multiple classifier consultation strategy to create an effective and powerful classifier for polyp identification. This strategy benefits from recent findings that different classification models can better learn and extract various information within the image. Therefore, our Ensemble classifier can derive a more consequential decision than each individual classifier. The extracted combined information inherits the ResNet's advantage of residual connection, while it also extracts objects when covered by occlusions through depth-wise separable convolution layer of the Xception model. Here, we applied our strategy to still frames extracted from a colonoscopy video. It outperformed other state-of-the-art techniques with a performance measure greater than 95% in each of the algorithm parameters. Our method will help researchers and gastroenterologists develop clinically applicable, computational-guided tools for colonoscopy screening. It may be extended to other clinical diagnoses that rely on image.

Keywords: colorectal cancer, deep learning, polyp detection, colonoscopy, ensemble classifier

1 INTRODUCTION

Cancer is a complex disease caused by uncontrolled cell growth. Colorectal cancer (CRC) is a form of cancer that occurs when irregular growth occurs in the colon and rectum (the last part of the gastrointestinal (GI) system). A polyp's initial stage is noncancerous; however, some polyps may become cancerous over time. For the determination of the treatment plan, the identification of a polyp is essential. Regular screening can prevent cancer through the identification and removal of precancerous polyps (Soh et al., 2018; Sánchez-Peralta et al., 2020). Diagnosis of the disease at an early stage can result in more effective treatment. As a consequence, screening decreases CRC mortality by both reducing the incidence and increasing survival. The visual test is a commonly recommended technique for CRC screening. Colonoscopy is one of the standard screening techniques for visualizing specific parts of the colon (Sánchez-Peralta et al., 2020). During a colonoscopy, gastroenterologists perform visual screening of the entire colon from the rectum to the cecum with the help of a light and tiny camera attached to the colonoscope.

Most of the works available in literature have focused on the detection of different types of polyps, such as cancerous or noncancerous, due to the lack of availability of a benchmark dataset. However, a colonoscopy video contains frames with polyps and without polyps. Therefore, as the first step, it is necessary to conduct a study to classify the frames to examine the presence of polyps, which will further study the features of the polyps, such as whether it is cancerous or not, location on the colorectum, or the disease stages.

Multiple computer-aided design approaches have been proposed in previous studies that can be applied to CRC analysis. In this direction, most of the works have used k-means, Fuzzy C-means, K-Nearest Neighbor (KNN), and support vector machine (SVM) based on handcrafted features (Häfner et al., 2015; Wimmer et al., 2016; Ševo et al., 2016; Shin and Balasingham, 2017; Sanchez-Gonzalez et al., 2018; Sundaram and Santhiyakumari, 2019). For example, Oh et al. (2007) used edge detection–based methods and achieved 96.5% accuracy in detecting informative frames. Recent studies have introduced the applicability of deep learning in colon cancer detection (Bernal et al., 2017; Pacal et al., 2020). Bernal et al. (2017) compared the efficacy of handcrafted features with CNN-extracted features in detecting polyp presence on still frames. They claimed that end-to-end learning approaches based on the CNN are more efficient than those based on handmade features. Akbari et al., (2018) applied the CNN on whole-slide images to classify informative and noninformative frames. Others (Ribeiro et al., 2016; Sharma et al., 2020a; Sharma et al., 2020b) also utilized deep learning architecture, such as VGG, ResNet, and GoogLeNet, for informative frame detection. Graham et al. (2019) used a minimum information loss deep neural network to segment the polyp region; they could achieve an F1 score of 0.825 and object-level dice score of 0.875. Sornapudi et al. (2019) proposed a CNN-based approach and used transfer learning from the ImageNet dataset to achieve an 88.28% F1 score in polyp segmentation.

The literature proffers a clear trend to eventually replace handcrafted features and traditional ML techniques with end-to-end frameworks. It all enables significant improvement in colonoscopy image analysis, making it more automated and providing more reliable and precise polyp detection methods. This work proposed a fully automatic system to classify polyps on still-frames from colonoscopy. The proposed system is an ensemble of different CNN architectures. The system will provide a decision in two stages. First, the frames are assessed as informative (frames containing polyps) and uninformative (frames not containing polyps). Second, the same classification model is applied to predict informative frames as cancerous (frames containing cancerous polyps) or noncancerous (frames containing non-cancerous polyps). Ensemble learning is an approach where better efficiency is obtained by integrating the results into one high-quality classifier from multiple classification models. Our methodology also addresses the problems involved in the use of the CNN for classification with limited sample data by using pre-trained CNN on a large dataset of natural images (> 1 million) and fine-tuning (optimizing) them using a smaller medical image dataset (at the thousand level). The different CNNs in our Ensemble method allow extracting the image features on different semantic levels so that the distinctive and subtle variations between different image classes can be identified. The contribution of this work includes the following three parts:

- Development of an automatic polyp detection model from colonoscopy images. Our model will classify the colonoscopy frames as informative or uninformative and further classify informative frames as cancerous or noncancerous.
- Detection of polyps during colonoscopy screening through the multiple classifier consultation strategy to create an effective and strong classifier for polyp identification. After our literature review, we assessed that it is the first approach by an Ensemble of various significant learning models for colonoscopy frame analysis.
- The robustness of the proposed Ensemble classifier is demonstrated by applying it to a real-world clinical dataset and comparing its result with the publicly available benchmark dataset. A suitable statistical significance test is conducted to assess the significant difference in the performance of proposed methods with a single classifier.

2 MATERIALS AND METHODS

Conventional techniques for classification tasks rely on manually examined features. Optimal feature selection plays a vital role in the final outcome of the selected computer vision task. Identifying the best features for a target segmentation/classification algorithm is difficult due to less intergroup variability. The variability in the visual appearance of polyps and their background is much less compared to the object and its background in natural images. Therefore, algorithms that are efficient for computer vision task in natural images are not always an ideal approach to deal with the computer vision task in medical imaging. Deep learning is an active domain in the research area of medical image analysis as it has recently successfully overcome the challenges in image recognition on the ImageNet dataset (Deng et al., 2009). Hence, the application of the CNN, a deep learning approach in CRC analysis, is

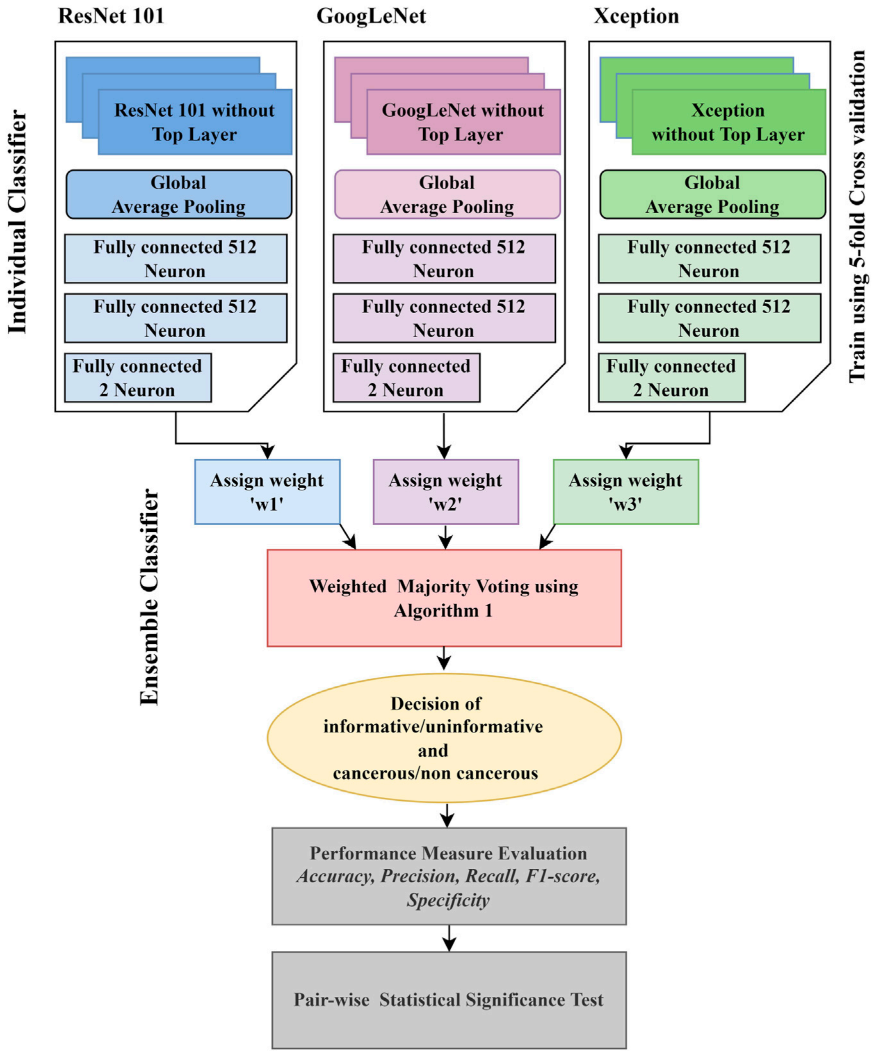

ResNet 101
GoogLeNet
Xception
Individual Classifier
ResNet 101 without Top Layer
GoogLeNet without Top Layer
Xception without Top Layer
Global Average Pooling
Fully connected 512 Neuron
Fully connected 512 Neuron
Fully connected 2 Neuron
Train using 5-fold Cross validation
Ensemble Classifier
Assign weight 'w1'
Assign weight 'w2'
Assign weight 'w3'
Weighted Majority Voting using Algorithm 1
Decision of informative/uninformative and cancerous/non cancerous
Performance Measure Evaluation
Accuracy, Precision, Recall, F1-score, Specificity
Pair-wise Statistical Significance Test

FIGURE 1 | Workflow of the proposed system.

TABLE 1 | Summary of datasets used in this study.

Dataset	# Frames		# Frames with polyps	
	Informative	Uninformative	Cancerous	Noncancerous
Aichi-Medical dataset	397	500	125	272
Kvasir dataset	500	500	-	-
Depeca colonoscopy dataset	—	—	55	21

introduced in this work. The workflow of the proposed work is described in **Figure 1**.

2.1 Dataset

A dataset that consists of colonoscopy frames extracted from a colonoscopy video is used in this work. The data were generated in the Department of Gastroenterology, Aichi Medical University, Nagakute, Japan, with the IRB approval of the Aichi Medical University ethical committee (15 January 2018; Approval No. 2017-H304). To assess the robustness of the proposed methodology, evaluation is performed on two publicly available benchmark datasets, Kvasir (Pogorelov et al., 2017) and Depeca (Mesejo et al., 2016). The mentioned datasets can be downloaded from https://datasets.simula.no/kvasir/and http://www.depeca.uah.es/colonoscopy_dataset/, respectively. The details of these datasets are summarized in **Table 1**. Because of the unavailability of any polyp dataset that contains only two groups, that is, cancerous and noncancerous, we combine serrated and adenoma frames available

TABLE 2 | Performance measures for evaluating the detection model.

Measures	Formula	Description
Accuracy Urban et al. (2018); Zhang et al. (2016); Bandyopadhyay et al. (2013); Bedrikovetski et al. (2021)	$\frac{\lvert TP\rvert+\lvert TN\rvert}{\lvert TP\rvert+\lvert TN\rvert+\lvert FP\rvert+\lvert FN\rvert}$	The ratio of the number of correct prediction with respect to total observations
Precision Urban et al. (2018); Zhang et al. (2016); Bandyopadhyay et al. (2013); Bedrikovetski et al. (2021)	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FP\rvert}$	The ratio of the number of correct positive prediction with respect to total positive prediction
Recall/Sensitivity Urban et al. (2018); Zhang et al. (2016); Bandyopadhyay et al. (2013); Bedrikovetski et al. (2021)	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}$	The ratio of number of correct positive prediction with respect to actual positive observation
F1 score/Dice-coefficient Urban et al. (2018); Zhang et al. (2016); Bandyopadhyay et al. (2013); Bedrikovetski et al. (2021)	$2 \times \frac{Recall \times Precision}{Recall+Precision}$	F1 score is the harmonic mean of both precision and recall

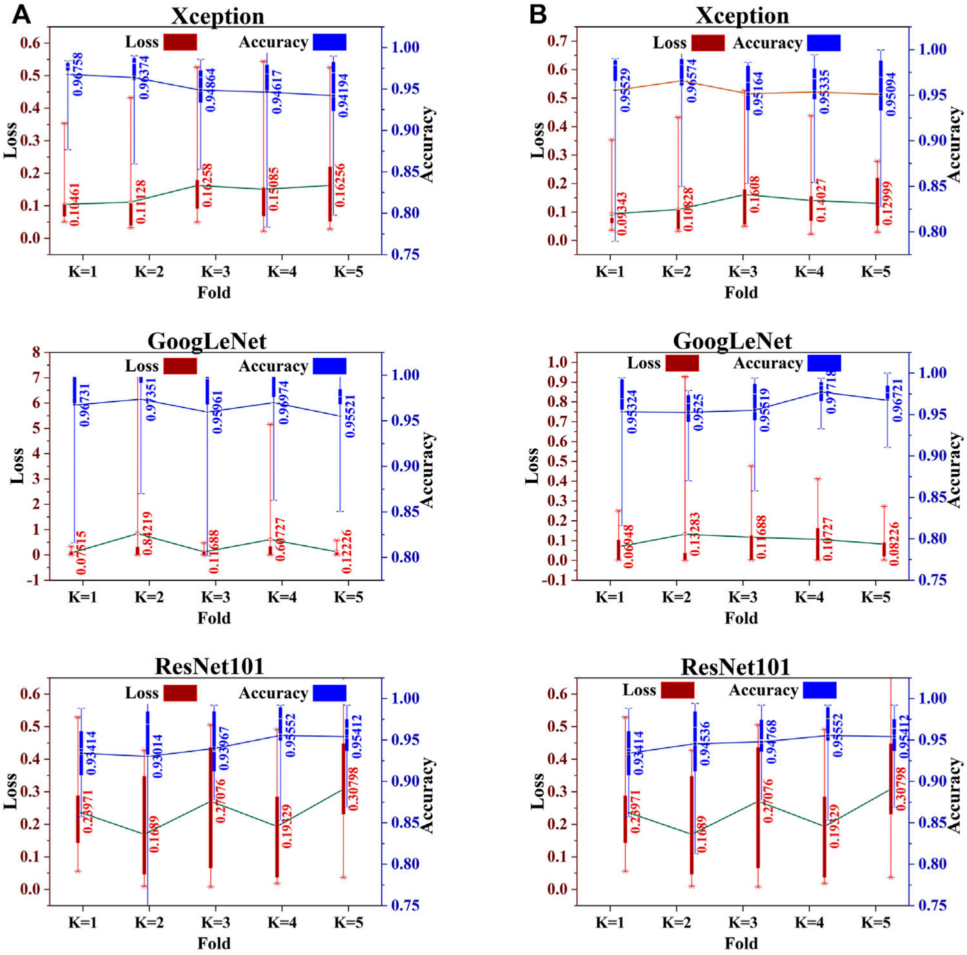


FIGURE 2 | Five-fold cross-validation accuracy and loss of each individual classifier for **(A)** informative frame detection and **(B)** cancerous and noncancerous polyp categorization.

on the Depeca colonoscopy dataset, which has a total of 55 original frames. We consider this combined class as cancerous and 21 hyperplastic frames as noncancerous in our study.

2.2 Classification Model

The CNN typically requires a massive dataset for training (at least thousands of samples if not available in millions). Thus, the application of the CNN trained from scratch is difficult because limited time and workload of experts to create labeled sample datasets on medical images. If the available training dataset is small in size, as is the case in this domain of medical image analysis, methods based on the CNN usually overfit and are unable to extract the image features in high quality.

Transfer learning is the strategy through which a CNN is initially trained to learn standardized image characteristics on a

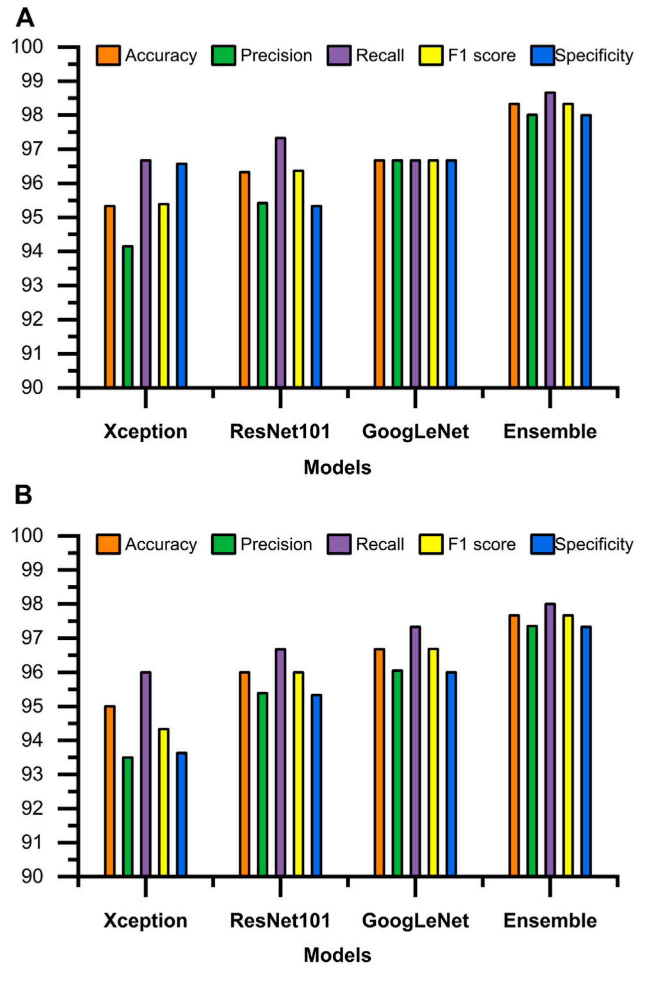


FIGURE 3 | Test results of all four classifiers for **(A)** informative frame detection and **(B)** cancerous and noncancerous polyp classification.

large-scale labeled image dataset and then used to retrieve similar features from a smaller dataset. It has already been successfully applied in different image analysis tasks or disease-related trials. Therefore, in our proposed Ensemble classifier, the base model weights are transfer-learned from the ImageNet dataset (Deng et al., 2009). Data augmentation is applied to all the datasets for balancing the dataset. The augmentation techniques such as shearing, rotation, skewing, zooming, and inverting are used. It is one of the most common approaches used for minimizing overfitting during the training phase of the CNN. This approach artificially expands the dataset using different class-preserving functions applied to each image to generate synthetic images. The concept behind the augmentation techniques is that the reproduced samples do not change their semantic meaning but enable the generation of a new sample to increase dataset size. As mentioned earlier, training CNNs on large data leads to improvement in its efficiency, robustness, and generalizability on previously unseen data or samples. Hence, in this work, we apply clock-wise rotation with an angle of 45°, 90°, and 120° and zooming parameters of 30.00 and 10.00% to the 1,000 original images of the Kvasir dataset to generate another 1,000 augmented images. Due to fewer data in the Aichi-Medical dataset, we apply a shearing operation with a value of 0.1 to original frames and the augmentation as mentioned above to balance the class disparity in the number of images. Again, for the Depeca colonoscopy dataset, we apply rotation, shearing, inverting, skewing, and zooming to obtain a total of 2000 images, including the original image. After applying augmentation, each individual dataset contains 2000 images. Then, a two-level classification is carried out in this research to fulfill the objective.

- The first-level classification is for informative frame detection. The outcome of the classifier is expected to be the class label of individual frames as informative or uninformative.
- The second classification is to detect cancerous polyps from informative frames. The outcome of the classifier is expected to be the class label of an individual informative frame as a cancerous or noncancerous polyp.

Three CNN architectures are used along with the proposed Ensemble classifier. The description of individual classifiers is widely available in the literature.

- ResNet101: As the information from the input or the gradient calculated by the CNN passes through many layers, it sometimes vanishes in between the hidden layers and sometimes rinsed out by the time it hits the end or beginning of the network (Simonyan and Zisserman, 2014; Huang et al., 2017). This was solved using ResNet. Conventional neural networks forward the output information of a layer (e.g., *Lth*) as an input to the successive layers $(L+1)^{th}$. If X is the input to the *Lth* layer, then the input to the $L+1$st layer will be X', where X' can be represented as

$$X' = f(X). \tag{1}$$

Here, f is the series of different operations within the convolution block. ResNets have added a skip-connection that bypasses the nonlinear transformations with an identity function (Szegedy et al., 2015)

$$X' = f(X) + X. \tag{2}$$

In this structure, input images are convolved by a kernel of size 7 × 7 with a stride equal to two followed by max-pooling. The first residual block accepts the output of this pooling layer. It uses a residual connection that adds the output of the pooling layer with the output of the first residual block. The residual block is constituted of three subsequent convolution layers. The first and third convolution operations are 1 × 1 convolution. The first convolution mixes up all the local properties of the image pixels across all the channels, and the convolution layer with 3 × 3 kernel mixes up the spatial properties. The third convolution layer helps increase the number of channels. The residual connection does not have any attenuation or gradient multiplication with activation. So, it is a unity gradient. By virtue of a residual connection, the exact value of the gradient can propagate back to the input layer. Using this structure, it is possible to carry forward information to the end of the model, but it is

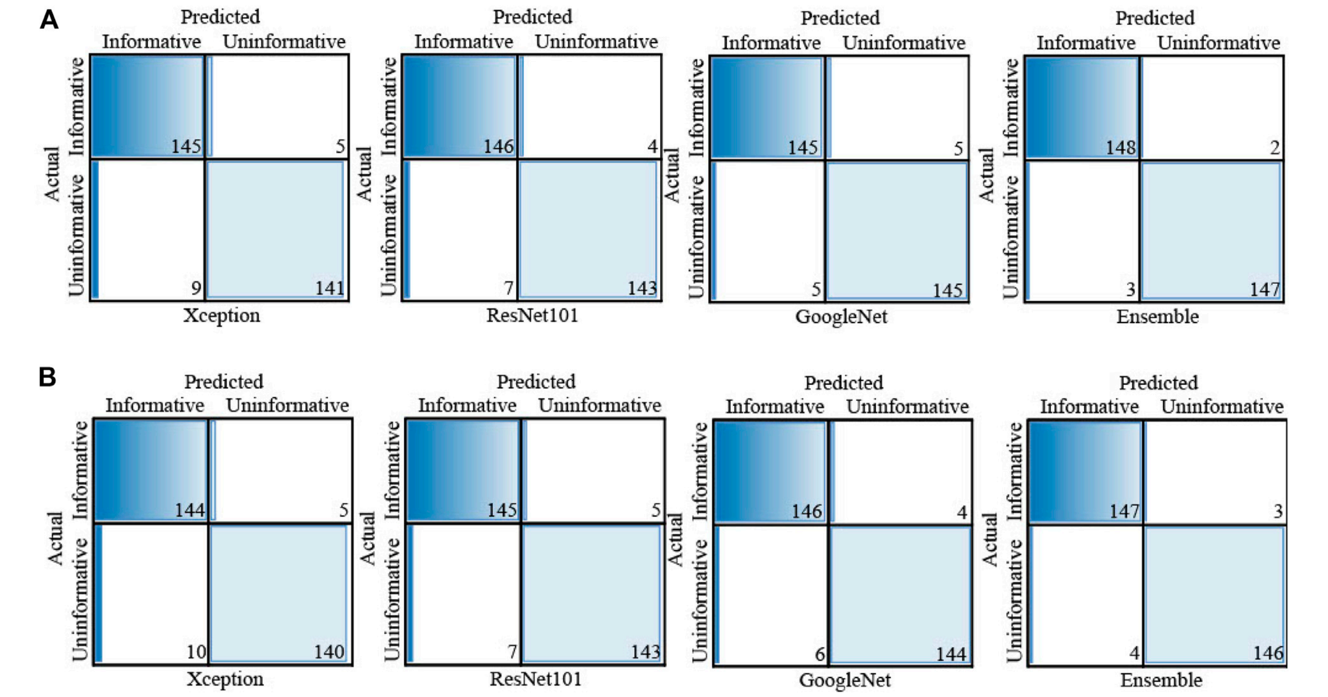


FIGURE 4 | Confusion matrix of each individual classifier for **(A)** informative frame detection and **(B)** cancerous and noncancerous polyp classification.

possible to backpropagate the gradient without vanishing it. The main power of ResNet is the direct flow of gradient through the identity relation from the successive layers to the prior layers.

- GoogLeNet: In the GoogLeNet architecture, a new "Inception" subnetwork module is added. The findings of various parallel convolution filters present at the inception are concatenated. The repetition of the Inception modules captures the optimal sparse representation of the image, while simultaneously reducing dimensionality. The network comprises 22 layers that require training (or 27 if pooling layers). Experiments have shown that GoogLeNet has fewer trainable weights than AlexNet and, thus, is more accurate (Szegedy et al., 2015).
- Xception: In the structure of Xception, the convolution layer used in ResNet is replaced by a depth-wise separable convolution module. Depth-wise separable convolution converges the process faster, and the accuracy is high. In the depth-wise separable convolution module, depth-wise convolution is followed by a 1x1 convolution. The number of filters is equal to the number of channels in each layer. With decreasing number of channels, the number of connections also decrease, which eliminates the drawback of performing convolution across all the channels. The depth-wise separable convolution learns spatial correlation, and the 1x1 convolution learns the interchannel correlation. The nonlinear activation function is not used. As state-of-the-art literature conveys that Xception outperforms VGG-16 and ResNet-152 in the ImageNet classification challenge (Chollet, 2017), Xception retains the characteristics of ResNet and can effectively deal with the complex situation of extracting targets covered by occlusions. Considering these advantages, in our proposed Ensemble method, we used Xception as a candidate model that is optimized based on ResNet.

Each individual classifier is fine-tuned according to our objective. Because the classification task in this work is to deal with binary classification problems, the models are fine-tuned by truncating the top layers of each model and replacing them with a modified fully connected network with a two-neuron output layer. Finding the best model for a specific task is dependent on efficient hyperparameter optimization. The best hyperparameters considered in this work are Adam as an optimizer, 0.001 learning rate, and a batch size of 32.

2.3 Ensemble Classifier

Ensemble classification is the preference of many scientists in a variety of fields such as computer vision and medical image analysis. For example, Bolón-Canedo et al. (2012) developed an Ensemble classification approach in the bioinformatics field, aiming for interpretation of the microarray data classification. Sun et al. (2015) implemented the concept of Ensemble classification on an imbalanced dataset. They reported that it outperformed conventional classification techniques. Sharif et al. (2020) applied an Ensemble classifier to analyze the data for squamous cell carcinoma. Bose et al. (2021) proposed an Ensemble classifier for efficient classification of a malignant tumor. Shakeel et al. (2020) used an Ensemble classifier to detect non–small cell lung cancer from CT images. Hussain et al. (2020) also used the Ensemble classifier to mine the data in cervical precancerous samples and cancer lesions. Some other biomedical research contributions based on the application of the Ensemble classifier to improve computer-aided systems can be found in references (Yang et al., 2016; Yang et al., 2020; Ayaz et al., 2021; Mahfouz et al., 2021). These ensembles are basically combining traditional machine learning models, such as SVM and AdaBoost, with one of the deep learning models. Our motivation for this work is to find a novel and efficient model for classifying colonic polyps to detect colorectal cancer. So far, there has been limited work that focuses on improving the performance of polyp detection using Ensemble. This encouraged us to incorporate the principle of Ensemble classification in this work. In the Ensemble method, the approach is to consult as many classifiers as possible and factor their decision in such a way that its efficiency will be enhanced. **Figure 1** presents an overview of the proposed Ensemble method. Unlike most other ensembles in the literature, which rely on handcrafted features, we use three of the best performing CNN models in both the computer vision and medical imaging tasks in our Ensemble. Initially, the CNN architectures whose weights have been initialized on natural image data are fine-tuned. Each of the fine-tuned CNN extracts independent image features to classify an image. Then, the Ensemble classifier chooses the class label for a particular image based on the decision of each candidate classifier. To consider the decision of each individual classifier, a weight is assigned to each individual decision based on the weighted majority voting technique. The process of decision making by the Ensemble classifier is detailed in Algorithm 1. During the decision-making process, we consider the loss of each individual model when deciding the class-label probability for each image. The individual model that has the smallest loss will be assigned the highest weight.

Algorithm 1. Decision of the Ensemble classifier.

```
Input: Input frames as still images ($I_i$).
Output: Class label for each image.
Initialization: matrix[M][P]; N=Total number of images in a dataset; M=3, M is the number of classifiers;
P=2, P is the number of classes.
Algorithm:
for $i$ = 0 to $N$ do
    temp = 0
    $Z_{max}$ = 0
    for $j$ = 0 to $M$ do
        for $k$ =0 to $P$ do
            if $X_j$ select $x_k$ for $I_i$ then
                $\delta_k(x, y)$ = 1          ▷ if classifier j chooses class k from P, then $\delta_k = 1$ otherwise, $\delta_k = 0$
            else
                $\delta_k(x, y)$ = 0
            end if
            $matrix_{jk}$ =$\delta_k(x, y)$
            $W_k$ = log$\frac{1}{E_k}$          ▷ $E_k$ is the loss and $W_k$ is the assigned weight for choosing class k
            $Z_j$ = $Z_j + W_k \times \delta_k(x, y)$
        end for
        $Z_{max}$ = $Z_j$          ▷ gives the j index to find the classifier M that holds max value
        if $Z_{max}$ = $Z_j$ then
            temp = j
        end if          ▷ find the class index with corresponding matrix value 1
    end for
    for $l$ = 0 to $M$ do
        if $matrix_{temp,l}$ = 1 then
            return $l$          ▷ return class with maximum value
        end if
    end for
end for
```

2.4 Performance Metrics for Evaluation of Classification Task

The performance evaluation parameters of a classification model are based on the correct and incorrect estimation of test records anticipated by the model. The confusion matrix gives the insight of predicated values compared to the actual values that can be visualized for the test dataset for all the classes. The four measures, true positive (TP), false positive (FP), true negative (TN), and false negative (FN), are part of the confusion matrix. Based on these four measures, efficient parameters to evaluate different classification techniques can be estimated. The most common performance measures based on the confusion matrix are explained in **Table 2**. This work has considered accuracy, precision, recall, F1 score, and specificity to evaluate the performance of our Ensemble classifier.

2.5 Statistical Analysis

The statistical significance test is applied to compare the significance of our proposed Ensemble method with others. We used the McNemar test (McNemar, 1947; Demšar, 2006) with a contingency table. The McNemar test is used to compare the accuracy of prediction for two models.

3 RESULTS AND DISCUSSION

To evaluate the efficiency of the proposed method and compare their performance with the existing methods, we applied and evaluated the proposed Ensemble method along with the individual classifier on our generated dataset. These models were implemented using the Keras deep learning framework with a TensorFlow backend provided by Google-Colab.

The dataset was split into two subsets using the train–test strategy. We first consider splitting with a ratio of 0.15. The first subset of 300 images is considered only for testing model performance, while the second subset of 1700 images is used for training. In the training phase, five-fold cross-validation is applied wherein each fold with 15% of the training data is considered for validation to improve the performance of the model. We train the same base classifier individually for each classification task to achieve both the objectives of this work. First, we perform the classification of informative and uninformative frames. Then, we conduct a separate training for all three classifiers for the second classification purpose, that is, to classify cancerous and noncancerous polyps. The box and whisker plots in **Figures 2A,B** show the mean score of the validation accuracy and loss achieved during each fold for each individual classifier. GoogLeNet achieved 96.5% average accuracy after five-fold cross-validation, which is the highest among all three individual classifiers for both classification tasks. For the test dataset, the accuracy, precision, recall, F1 score, and specificity values were reported for all the classifiers.

3.1 Evaluation of Classifier Performance

Figures 3, **4** display the performance of the Ensemble classifier along with each individual classifier on the generated dataset. For informative frame detection, our proposed Ensemble obtained 98.3, 98.6, and 98.01% accuracy, precision, and recall, respectively, and for cancerous polyp detection, 97.66, 98.66, and 96.73% accuracy, precision, and recall, respectively. We performed receiver operating characteristic (ROC) analysis, and **Figure 5** shows the model performance using the measured area under the ROC curve (AUC). These observed

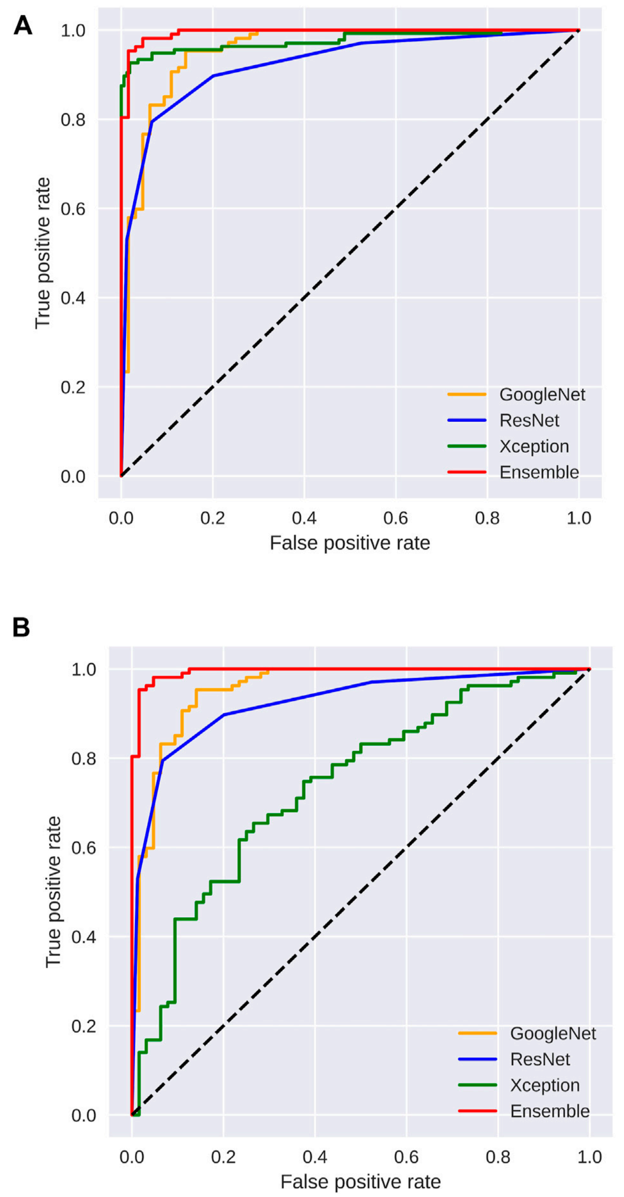


FIGURE 5 | Area under the ROC curve analysis for **(A)** informative frame detection and **(B)** cancerous and noncancerous polyp classification.

results indicated that our proposed Ensemble classifier performed better than any other classifiers for both classification tasks.

Based on our objective of this work, both the FP and FN are crucial, and our goal is to keep them low. In the first scenario, our proposed system informs that patients having a cancerous tumor but being labeled as noncancerous could lead to misclassification denoted as false negative (FN). In another scenario, patients not having a cancerous tumor but being informed as abnormal (cancerous) could cause false positive (FP). Both FNs and FPs have a significant impact on misclassification, therefore leading to wrong diagnosis and causing human health problems. We considered F1 score along with other performance evaluation measures to equally prioritize both FP and FN. We observed that our proposed method gives the highest F1 score of 98% for informative frame detection and 97.33% for cancerous polyp detection. Almost equal precision, recall, and F1 score of our ensemble convey that the proposed model has a negligible rate of misclassification, which is also supported by the specificity value.

Figure 6A shows the comparison of our Ensemble classifier's result on the Kvasir dataset with our dataset. The Kvasir dataset is considered a benchmark dataset for informative frame detection, and our Ensemble attains a value of test accuracy 98%, precision 99.33%, recall 96.75%, F1 score 98.03%, and specificity 99.31%. **Figure 6B** compares the Ensemble classifier's result on the Depeca colonoscopy dataset to produce the effectiveness of our proposed method on a new independent dataset for

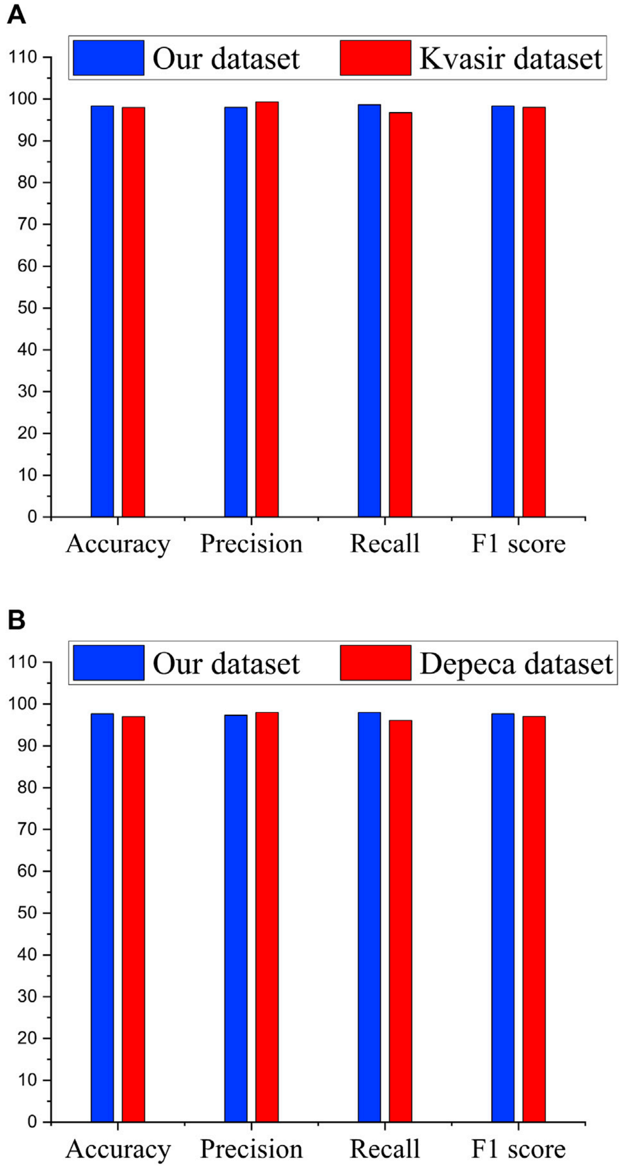


FIGURE 6 | Performance comparison of Ensemble classifiers. **(A)** Performance of Ensemble classifier on significant frame detection. **(B)** Performance of Ensemble classifier on classification of cancerous and noncancerous polyps.

TABLE 3 | Classification performance in comparison with similar work.

Objective	Methods	Algorithm	Accuracy	Precision	Recall	F1 Score	Specificity
Informative frame detection	Proposed Ensemble	CNN	**98.3**	98.6	**98.01**	**98.33**	98.66
	Akbari et al. (2018)	CNN	90.28	74.34	68.32	71.20	94.97
	Zhang et al. (2016)	Ensemble (SVM + CNN)	98.0	**99.4**	97.6	98.00	-
	Shin and Balasingham (2017)	CNN	86.69	86.28	28.90	43.30	99.02
	Liew et al. (2021)	Ensemble (ResNet50 + Adaboost)	97.91	99.35	96.45	—	**99.38**
Cancerous and noncancerous polyp identification	Proposed Ensemble	CNN	**97.66**	**98.66**	**96.73**	**97.68**	**98.63**
	Urban et al. (2018)	CNN	90.00	—	88.1	—	—
	Zhang et al. (2016)	Ensemble (SVM + CNN)	85.90	87.30	87.60	87.00	—
	Wang et al. (2018)	CNN	90.00	—	94.50	—	—
	Patino-Barrientos et al. (2020)	CNN	83.00	81.00	86.00	83.00	—

Bold values indicate the best performance.

TABLE 4 | Significance of Ensemble classifier decision in comparison with individual classifiers.

Classifier	Chi-squared Value	*p*-Value*
ResNet101 vs. Ensemble	4.16	0.041
GoogLeNet vs. Ensemble	2.28	0.039
Xception vs. Ensemble	6.75	0.009
ReNet101 vs. Ensemble	2.25	0.033
GoogLeNet vs. Ensemble	2.25	0.033
Xception vs. Ensemble	5.81	0.015

**p-value is based on the McNemar test.*

cancerous polyp detection. We observed that the results were consistent on all the datasets, which provides clear evidence of the robustness of our proposed method.

In **Table 3**, we compared the Ensemble classifier with other classifiers found in existing literature for CRC detection (Zhang et al., 2016; Shin and Balasingham, 2017; Akbari et al., 2018; Urban et al., 2018; Wang et al., 2018; Sharma et al., 2020a). From the observation, it is comprehendible that the proposed method outperforms the other classifier and is efficient in fulfilling our objective.

In aid of this interpretation, the McNemar test result of our Ensemble paired with each individual classifier is summarized in **Table 4**. The *p*-values obtained from this test are less than 0.05 in all the cases. These results show that our proposed Ensemble approach is superior to the other classifiers.

3.2 Discussion

Based on the results, it is confirmed that the proposed Ensemble classifier is an efficient model for colonoscopy image analysis and can be used as an assistant tool by the gastroenterologist during the screening of CRC. Even though the proposed method shows better performance, some clinical information such as sex and age of the patients, other medical conditions, geographic location, etc. are not considered in this work. Future work conducted by considering these criteria can improve computer-aided systems for early cancer detection and treatment in personalized medicine. As the massive dataset available for transfer learning contains natural images, the transfer-learned features are more reflective of the natural image characteristics and may not always necessarily reflect the subtle characteristics of medical images. Therefore, it is expected that transfer learning from the same domain large-scale dataset will lead to developing a more efficient automatic system for CRC analysis. Kudo et al. (1996) has reported that the detection rate to differentiate cancerous and noncancerous lesions using images from magnifying endoscopy is higher (81.5%) than that of the stereomicroscopic analysis. Therefore, a performance comparison of the proposed model considering the images of magnifying endoscopy and the colonoscopy images will be a future direction.

4 CONCLUSION

In this article, we introduced a new Ensemble method for the classification of each individual frame of a colonoscopy video as informative or uninformative and then for predicting the classified informative frames as cancerous or noncancerous polyps. Our Ensemble uses multiple fine-tuned CNNs that can learn diverse information present in individual images. The Ensemble can fuse the fine-tuned CNN models to derive a more powerful image classification scheme than the individual CNNs. When Xception extracts features, it achieves the best performance because Xception is optimized on the basis of ResNet, which makes Xception inherit not only ResNet's advantage of residual connection but also its ability to extract objects when covered by occlusions through depth-wise separable convolution. The analysis by the McNemar statistical test indicates high significance in the performance of the Ensemble classifier when compared to the individual classifiers. Therefore, our Ensemble shows the best performance for polyp detection on colonoscopy with an acceptable level of all performance measures in the range 0.95–1. A minor difference in precision and recall value of our Ensemble classifier indicates that it can accurately detect the presence of a polyp and also differentiate the cancerous from noncancerous polyps efficiently.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available due to ethical restrictions; we cannot publish the data currently. If required, we can provide the sample dataset after acceptance. Requests to access the datasets should be directed to KB, kangkana.bora@cottonuniversity.ac.in.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Aichi Medical University Ethical Committee Approval No. 2017-H304. 15 January 2018. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PS and BB conceived the study and made the study design. KK collected the data, and KK and KB designed the dataset and ground truth. PS, KB, and SM performed the statistical analysis. PS, KB, SM, BB, and ZZ participated in result interpretation and manuscript writing. PS, BB, SM, and ZZ wrote and edited the manuscript. All the authors read and approved the final manuscript.

FUNDING

ZZ was partially supported by the Cancer Prevention and Research Institute of Texas (CPRIT RP180734) and the Precision Health Chair Professorship fund. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the entire manuscript.

ACKNOWLEDGMENTS

We thank all the members of the Bioinformatics and Systems Medicine Laboratory and Computer Vision and Machine Learning Lab, NIT Meghalaya for the helpful suggestion.

REFERENCES

Akbari, M., Mohrekesh, M., Rafiei, S., Reza Soroushmehr, S. M., Karimi, N., Samavi, S., et al. (2018). Classification of Informative Frames in Colonoscopy Videos Using Convolutional Neural Networks with Binarized Weights. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2018, 65–68. doi:10.1109/EMBC.2018.8512226

Ayaz, M., Shaukat, F., and Raja, G. (2021). Ensemble Learning Based Automatic Detection of Tuberculosis in Chest X-ray Images Using Hybrid Feature Descriptors. *Phys. Eng. Sci. Med.* 44, 183–194. doi:10.1007/s13246-020-00966-0

Bandyopadhyay, S., Mallik, S., and Mukhopadhyay, A. (2013). A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 11, 95–115.

Bedrikovetski, S., Dudi-Venkata, N. N., Maicas, G., Kroon, H. M., Seow, W., Carneiro, G., et al. (2021). Artificial Intelligence for the Diagnosis of Lymph Node Metastases in Patients with Abdominopelvic Malignancy: A Systematic Review and Meta-Analysis. *Artif. Intelligence Med.* 113, 102022. doi:10.1016/j.artmed.2021.102022

Bernal, J., Tajkbaksh, N., Sanchez, F. J., Matuszewski, B. J., Chen, H., Yu, L., et al. (2017). Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the Miccai 2015 Endoscopic Vision challenge. *IEEE Trans. Med. Imaging* 36, 1231–1249. doi:10.1109/tmi.2017.2664042

Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2012). An Ensemble of Filters and Classifiers for Microarray Data Classification. *Pattern Recognition* 45, 531–539.

Chollet, F. (2017). "Xception: Deep Learning with Depthwise Separable Convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1251–1258. doi:10.1109/cvpr.2017.195

Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Machine Learn. Res.* 7, 1–30.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A Large-Scale Hierarchical Image Database," in 2009 IEEE conference on computer vision and pattern recognition (Miami, FL, USA: IEEE), 248–255. doi:10.1109/cvpr.2009.5206848

Graham, S., Chen, H., Gamper, J., Dou, Q., Heng, P.-A., Snead, D., et al. (2019). Mild-net: Minimal Information Loss Dilated Network for Gland Instance Segmentation in colon Histology Images. *Med. image Anal.* 52, 199–211. doi:10.1016/j.media.2018.12.001

Häfner, M., Tamaki, T., Tanaka, S., Uhl, A., Wimmer, G., and Yoshida, S. (2015). Local Fractal Dimension Based Approaches for Colonic Polyp Classification. *Med. Image Anal.* 26, 92–107. doi:10.1016/j.media.2015.08.007

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely Connected Convolutional Networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708. doi:10.1109/cvpr.2017.243

Hussain, E., Mahanta, L. B., Das, C. R., and Talukdar, R. K. (2020). A Comprehensive Study on the Multi-Class Cervical Cancer Diagnostic Prediction on Pap Smear Images Using a Fusion-Based Decision from Ensemble Deep Convolutional Neural Network. *Tissue and Cell* 65, 101347. doi:10.1016/j.tice.2020.101347

Kudo, S.-E., Tamura, S., Nakajima, T., Yamano, H.-o., Kusaka, H., and Watanabe, H. (1996). Diagnosis of Colorectal Tumorous Lesions by Magnifying Endoscopy. *Gastrointest. Endosc.* 44, 8–14. doi:10.1016/s0016-5107(96)70222-5

Liew, W. S., Tang, T. B., Lin, C.-H., and Lu, C.-K. (2021). Automatic Colonic Polyp Detection Using Integration of Modified Deep Residual Convolutional Neural Network and Ensemble Learning Approaches. *Comput. Methods Programs Biomed.* 206, 106114. doi:10.1016/j.cmpb.2021.106114

Mahfouz, M. A., Shoukry, A., and Ismail, M. A. (2021). Eknn: Ensemble Classifier Incorporating Connectivity and Density into Knn with Application to Cancer Diagnosis. *Artif. Intelligence Med.* 111, 101985. doi:10.1016/j.artmed.2020.101985

McNemar, Q. (1947). Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* 12, 153–157. doi:10.1007/bf02295996

Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., et al. (2016). Computer-aided Classification of Gastrointestinal Lesions in Regular Colonoscopy. *IEEE Trans. Med. Imaging* 35, 2051–2063. doi:10.1109/tmi.2016.2547947

Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., and de Groen, P. C. (2007). Informative Frame Classification for Endoscopy Video. *Med. Image Anal.* 11, 110–127. doi:10.1016/j.media.2006.10.003

Pacal, I., Karaboga, D., Basturk, A., Akay, B., and Nalbantoglu, U. (2020). A Comprehensive Review of Deep Learning in colon Cancer. *Comput. Biol. Med.* 126, 104003. doi:10.1016/j.compbiomed.2020.104003

Patino-Barrientos, S., Sierra-Sosa, D., Garcia-Zapirain, B., Castillo-Olea, C., and Elmaghraby, A. (2020). Kudo's Classification for Colon Polyps Assessment Using a Deep Learning Approach. *Appl. Sci.* 10, 501. doi:10.3390/app10020501

Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., et al. (2017). "Kvasir: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in Proceedings of the 8th ACM on Multimedia Systems Conference, 164–169.

Ribeiro, E., Uhl, A., Wimmer, G., and Häfner, M. (2016). Exploring Deep Learning and Transfer Learning for Colonic Polyp Classification. *Comput. Math. Methods Med.* 2016, 6584725. doi:10.1155/2016/6584725

Sánchez-González, A., García-Zapirain, B., Sierra-Sosa, D., and Elmaghraby, A. (2018). Automatized colon Polyp Segmentation via Contour Region Analysis. *Comput. Biol. Med.* 100, 152–164. doi:10.1016/j.compbiomed.2018.07.002

Sánchez-Peralta, L. F., Bote-Curiel, L., Picón, A., Sánchez-Margallo, F. M., and Pagador, J. B. (2020). Deep Learning to Find Colorectal Polyps in Colonoscopy: A Systematic Literature Review. *Artif. intelligence Med.* 2020, 101923.

Ševo, I., Avramović, A., Balasingham, I., Elle, O. J., Bergsland, J., and Aabakken, L. (2016). Edge Density Based Automatic Detection of Inflammation in Colonoscopy Videos. *Comput. Biol. Med.* 72, 138–150. doi:10.1016/j.compbiomed.2016.03.017

Shakeel, P. M., Burhanuddin, M., and Desa, M. I. (2020). Automatic Lung Cancer Detection from Ct Image Using Improved Deep Neural Network and Ensemble Classifier. *Neural Comput. Appl.* 2020, 1–14. doi:10.1007/s00521-020-04842-6

Shanmuga Sundaram, P., and Santhiyakumari, N. (2019). An Enhancement of Computer Aided Approach for colon Cancer Detection in Wce Images Using Roi Based Color Histogram and Svm2. *J. Med. Syst.* 43, 29. doi:10.1007/s10916-018-1153-9

Sharif, M. S., Abbod, M., Al-Bayatti, A., Amira, A., Alfakeeh, A. S., and Sanghera, B. (2020). An Accurate Ensemble Classifier for Medical Volume Analysis: Phantom and Clinical Pet Study. *IEEE Access* 8, 37482–37494. doi:10.1109/access.2020.2975135

Sharma, P., Bora, K., and Balabantaray, B. K. (2020a). "Identification of Significant Frames from Colonoscopy Video: An Approach Towardsearly Detection of Colorectal Cancer," in 2020 International Conference on Computational Performance Evaluation (ComPE) (Shillong, Meghalaya: IEEE), 316–320. doi:10.1109/compe49325.2020.9200003

Sharma, P., Bora, K., Kasugai, K., and Kumar Balabantaray, B. (2020b). Two Stage Classification with Cnn for Colorectal Cancer Detection. *ONCOLOGIE* 22, 129–145. doi:10.32604/oncologie.2020.013870

Simonyan, K., and Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.

Soh, N. Y. T., Chia, D. K. A., Teo, N. Z., Ong, C. J. M., and Wijaya, R. (2018). Prevalence of Colorectal Cancer in Acute Uncomplicated Diverticulitis and the Role of the Interval Colonoscopy. *Int. J. Colorectal Dis.* 33, 991–994. doi:10.1007/s00384-018-3039-1

Sornapudi, S., Meng, F., and Yi, S. (2019). Region-based Automated Localization of Colonoscopy and Wireless Capsule Endoscopy Polyps. *Appl. Sci.* 9, 2404. doi:10.3390/app9122404

Subash Chandra Bose, S., Sivanandam, N., and Praveen Sundar, P. V. (2021). Design of Ensemble Classifier Using Statistical Gradient and Dynamic Weight Logitboost for Malicious Tumor Detection. *J. Ambient Intell. Hum. Comput* 12, 6713–6723. doi:10.1007/s12652-020-02295-2

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., and Zhou, Y. (2015). A Novel Ensemble Method for Classifying Imbalanced Data. *Pattern Recognition* 48, 1623–1637. doi:10.1016/j.patcog.2014.11.014

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going Deeper with Convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9. doi:10.1109/cvpr.2015.7298594

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., et al. (2018). Deep Learning Localizes and Identifies Polyps in Real Time with 96% Accuracy in Screening Colonoscopy. *Gastroenterology* 155, 1069–e8. doi:10.1053/j.gastro.2018.06.037

Wang, P., Xiao, X., Glissen Brown, J. R., Berzin, T. M., Tu, M., Xiong, F., et al. (2018). Development and Validation of a Deep-Learning Algorithm for the Detection of Polyps during Colonoscopy. *Nat. Biomed. Eng.* 2, 741–748. doi:10.1038/s41551-018-0301-3

Wimmer, G., Tamaki, T., Tischendorf, J. J. W., Häfner, M., Yoshida, S., Tanaka, S., et al. (2016). Directional Wavelet Based Features for Colonic Polyp Classification. *Med. image Anal.* 31, 16–36. doi:10.1016/j.media.2016.02.001

Yang, C., Jiang, Z.-K., Liu, L.-H., and Zeng, M.-S. (2020). Pre-treatment Adc Image-Based Random forest Classifier for Identifying Resistant Rectal Adenocarcinoma to Neoadjuvant Chemoradiotherapy. *Int. J. Colorectal Dis.* 35, 101–107. doi:10.1007/s00384-019-03455-3

Yang, J.-J., Li, J., Shen, R., Zeng, Y., He, J., Bi, J., et al. (2016). Exploiting Ensemble Learning for Automatic Cataract Detection and Grading. *Comput. Methods Programs Biomed.* 124, 45–57. doi:10.1016/j.cmpb.2015.10.007

Younghak Shin, Y., and Balasingham, I. (2017). Comparison of Hand-Craft Feature Based Svm and Cnn Based Deep Learning Framework for Automatic Polyp Classification. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2017, 3277–3280. doi:10.1109/EMBC.2017.8037556

Zhang, R., Zheng, Y., Mak, T. W., Yu, R., Wong, S. H., Lau, J. Y., et al. (2016). Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level Cnn Features from Nonmedical Domain. *IEEE J. Biomed. Health Inform.* 21, 41–47. doi:10.1109/JBHI.2016.2635662

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

ORIGINAL RESEARCH
published: 18 May 2022
doi: 10.3389/fgene.2022.894024

Angiogenic Factor-Based Signature Predicts Prognosis and Immunotherapy Response in Non-Small-Cell Lung Cancer

Xinpei Gu[1†], *Liuxi Chu*[2†] *and Yanlan Kang*[3*]

[1]*Department of Human Anatomy, Shandong First Medical University and Shandong Academy of Medical Sciences, Taian, China,* [2]*School of Biological Sciences and Medical Engineering, Southeast University, Nanjing, China,* [3]*Institute of AI and Robotics, Academy for Engineering and Technology, Fudan University, Shanghai, China*

Non-small-cell lung cancer (NSCLC) is one of the most common malignancies, and specific molecular targets are still lacking. Angiogenesis plays a central regulatory role in the growth and metastasis of malignant tumors and angiogenic factors (AFs) are involved. Although there are many studies comparing AFs and cancer, a prognostic risk model for AFs and cancer in humans has not been reported in the literature. This study aimed to identify the key AFs closely related to the process of NSCLC development, and four genes have been found, C1QTNF6, SLC2A1, PTX3, and FSTL3. Then, we constructed a novel prognostic risk model based on these four genes in non-small-cell lung cancer (NSCLC) and fully analyzed the relationship with clinical features, immune infiltration, genomes, and predictors. This model had good discrimination and calibration and will perform well in predicting the prognosis of treatment in clinical practice.

Keywords: NSCLC, angiogenic factors, immunotherapy response, model validation, biomarkers

OPEN ACCESS

Edited by:
Saurav Mallik,
Harvard University, United States

Reviewed by:
Asim Bikas Das,
National Institute of Technology Warangal, India
Antonio Russo,
University of Palermo, Italy

****Correspondence:***
Yanlan Kang
1641866831@qq.com

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *11 March 2022*
Accepted: *14 April 2022*
Published: *18 May 2022*

Citation:
Gu X, Chu L and Kang Y (2022) Angiogenic Factor-Based Signature Predicts Prognosis and Immunotherapy Response in Non-Small-Cell Lung Cancer. Front. Genet. 13:894024. doi: 10.3389/fgene.2022.894024

1 INTRODUCTION

Lung cancer is one of the malignant tumors with the highest incidence and mortality worldwide (Hirsch et al., 2017). Every year, 1.8 million people (11.6% of total cases) are diagnosed with lung cancer, and about 1.6 million people (18.4% of total cancer deaths) died because of lung cancer. There are two basic forms of lung cancer, small cell lung cancer (SCLC) and non-small-cell lung cancer (NSCLC), and NSCLC accounts for approximately 85% (Ettinger et al., 2013; Gridelli et al., 2015). NSCLC is characterized by poor survival, and despite significant advances in new chemotherapeutic drugs and clinical surgery, the prognosis remains suboptimal (Ettinger et al., 2013; Gridelli et al., 2015; Ettinger et al., 2021). With the advent of targeted molecular therapy and immune checkpoint inhibitors, the use of biomarkers in identifying patients is becoming increasingly common (Ma et al., 2019; Wang et al., 2019). The existing evidence has suggested that targeted therapies have favorable therapeutic effects. However, acquired resistance has become a major obstacle in the field of targeted therapies (Chatterjee and Bivona 2019). Thus, more novel driver genes, therapeutic targets, and prognostic biomarkers must be discovered and used for targeted therapy in larger populations, more accurate prognosis prediction, and a better understanding of the mechanisms of lung cancer development.

Tumors can promote tumor angiogenesis, leading to angiogenesis, which is the one of hallmarks of cancer (Hanahan and Weinberg 2000). The process of new blood vessel formation is critical in supporting tumor growth, and solid tumors secrete angiogenic factors (AFs) implicated in the complex regulation of angiogenesis (Goveia et al., 2020). Numerous important target molecules of

TABLE 1 | Clinic pathological data of patients with NSCLC in this study.

Characteristic		Number
Age	<60	720
	≥60	221
Pathologic_M	M0	698
	M1	30
	MX	208
	NA	5
Pathologic_N	N0	600
	N1	213
	N2	106
	N3	7
	NX	14
	NA	1
Pathologic_T	T1	262
	T2	529
	T3	108
	T4	39
	TX	3
Clinical stage	I	476
	II	264
	III	159
	IV	31
	NA	11
Follow up status	Alive	570
	Dead	371

AFs in NSCLC and other cancers, such as vascular endothelial growth factor (VEGF) (Zhang 2015) and epidermal growth factor receptor (EGFR) (Oxnard et al., 2011), have all become clinical targets for antitumor angiogenesis. Antiangiogenic medications are increasingly used as anticancer drugs for first-line treatment. Moreover, since the introduction of the first humanized anti-VEGF monoclonal antibody, bevacizumab (Avastin), available in 2004 (Ferrara et al., 2005), there have been nearly 30 antiangiogenic drugs approved by the FDA (Lugano et al., 2020). AFs are also expected to be optimal therapeutic targets. Several significant global studies noted that angiogenesis inhibitors combined with immunotherapy can enhance the curative effect. There is increasing evidence that targeting angiogenesis improves the efficiency of cancer immunotherapy. A programmed cell death 1 (PD-1) inhibitor and camrelizumab (AiRuiKa™) can improve the treatment effect of chemotherapeutics in multiple types of cancers (Markham and Keam 2019). However, apatinib, a vascular endothelial growth factor receptor 2 (VEGFR2) tyrosine kinase inhibitor, has been shown to increase the infiltration of $CD8^+$ T cells, reduce the recruitment of tumor-associated macrophages, and improve the effect of PD-1 inhibitors (Zhao et al., 2019).

Despite many studies investigating the association between AFs and cancers, whether AFs can be used as biomarkers to predict the prognosis of NSCLC patients is still unknown. In our study, based on the machine algorithms and bioinformatics methods, AF-related risk score (AFRS) was established. Four key prognosis-related AFs, C1QTNF6, SLC2A1, PTX3, and FSTL3, were first screened using bioinformatics analysis of differentially expressed genes (DEGs). Then, we attempted to construct a new risk score model to predict NSCLC, and we further analyzed the clinical features, immune infiltration, genomes, and multiple predictors. To further validate the AF-related prognostic risk score model, we used external dataset validation. An overview of this study is shown in **Supplementary Figure S1**.

2 RESULTS

The expression profile data of NSCLC patients were downloaded from the UCSC database. The detailed clinical features of these patients are summarized in **Table 1**.

2.1 Differential Expression Analysis and Functional Enrichment Analysis of Non-Small-Cell Lung Cancer

We identified a total of 372 differentially expressed AF genes in cancer and normal samples (with a threshold of adj.P.Val<0.01 & |log (FC) |≥1) (**Figures 1A,B**). GO and KEGG functional enrichment analyses of the differentially expressed AF genes

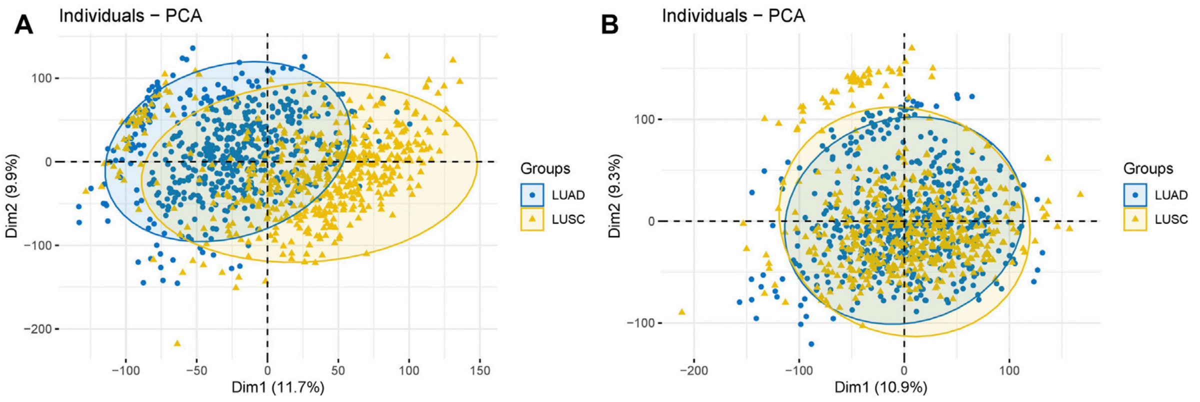


FIGURE 1 | Batch effect removal. **(A)** Before batch effects were removed. **(B)** After batch effects were removed.

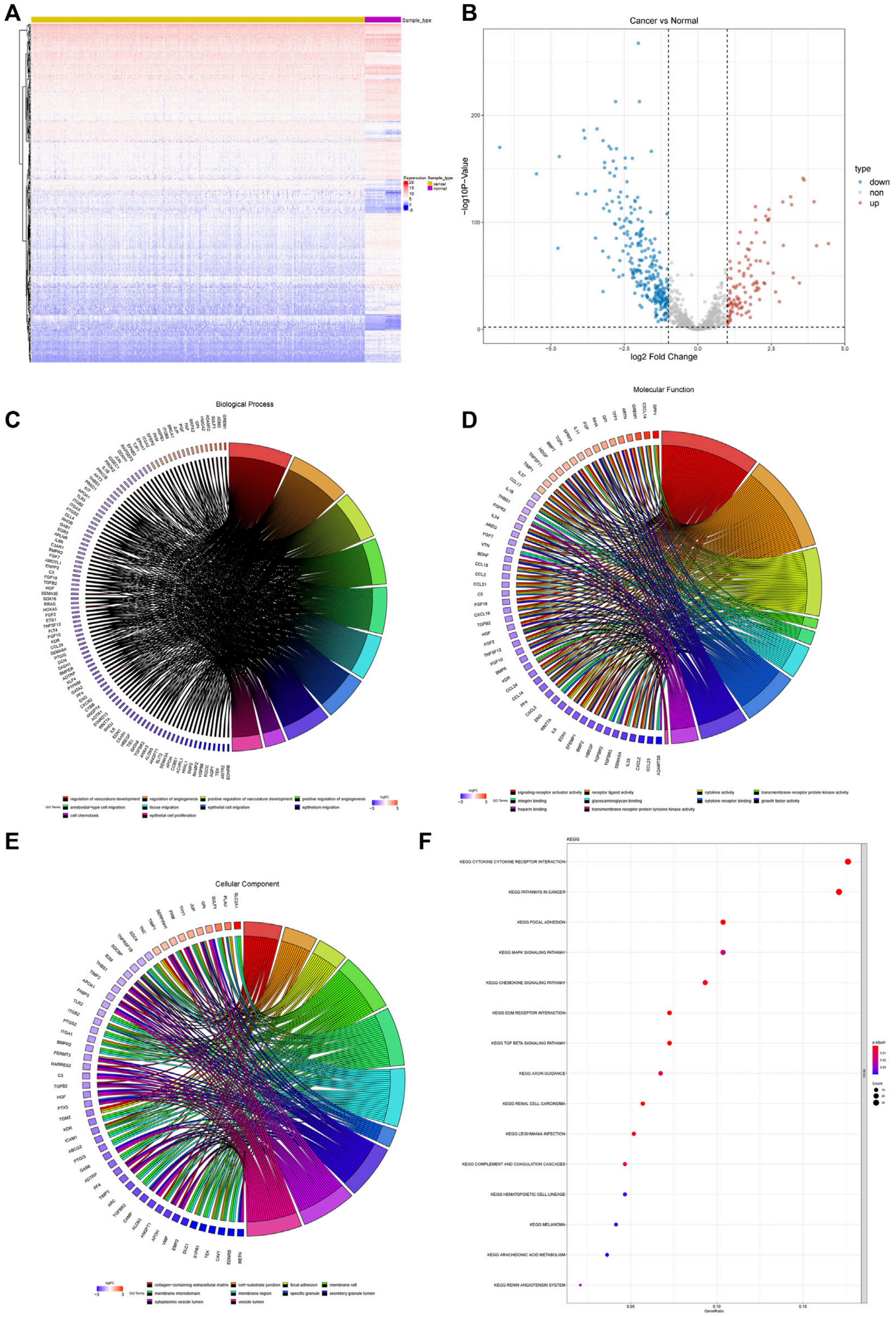


FIGURE 2 | Differential expression and functional enrichment of AF genes in non-small-cell lung cancer. **(A)** Heatmap and clustering of differentially expressed AF genes. **(B)** Volcano map of differentially expressed AF genes. **(C)** GO biological processes **(D)** GO molecular functions **(E)** GO cellular components and **(F)** KEGG.

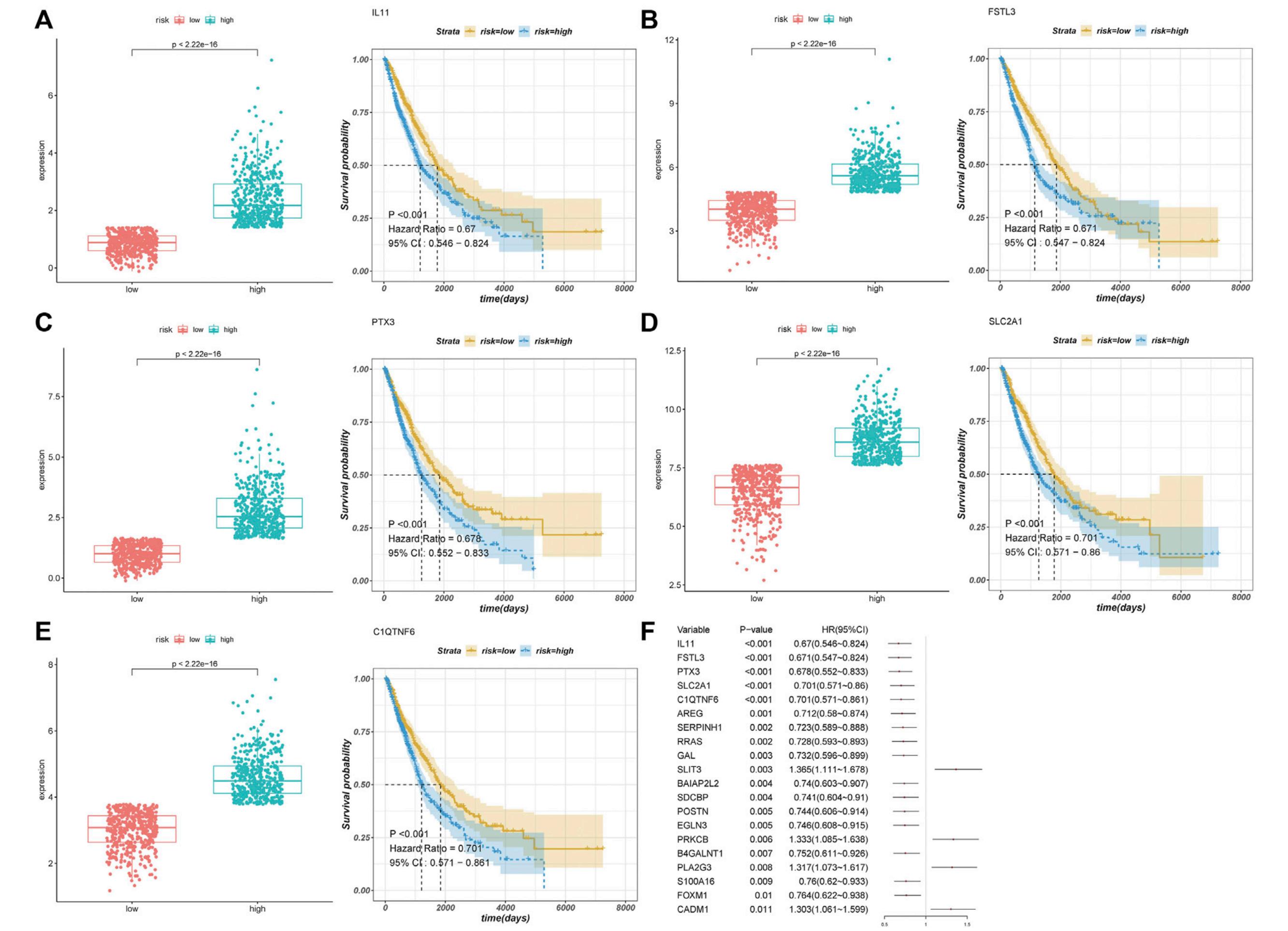


FIGURE 3 | Univariate Cox regression analysis. **(A–E)** Top five prognostic genes. **(F)** Forest plot of the top 20 genes.

were then performed (**Figures 1C–F**). The enriched GO terms of DEGs were classified into three categories: molecular functions, cellular components, and biological processes. The results revealed that these genes were enriched for GO terms such as regulation of vasculature development, regulation of angiogenesis, ameboidal-type cell migration, and positive regulation of vasculature development, epithelial cell proliferation, and tissue migration. The KEGG pathway enrichment showed the enrichment of critical pathways involved in tumorigenesis and metastasis, including pathways in cancer, focal adhesion, the MAPK signaling pathway, the chemokine signaling pathway, the TGF-β signaling pathway, and renal cell carcinoma. The top 15 highly enriched KEGG pathways are presented in **Figure 1F**.

2.2 Cox Regression Analysis of Differentially Expressed Angiogenic Factor Genes

We performed a univariate Cox regression analysis of these differentially expressed AF genes and identified 58 AF-related genes that were associated with the prognosis of NSCLC. We performed survival analyses of the top five genes in terms of the *p*-value (**Figures 2A–E**). The low expression of these five genes was associated with a worse prognosis (**Figure 2F**).

2.3 Development of Risk Model Using Lasso Regression

A total of five AF genes significantly associated with prognosis ($p < 0.001$) in the univariate Cox regression were further selected for lasso regression (**Supplementary Figure S2**). We first used cross-validation to identify the minimal lambda, i.e. lambda min, and then selected the four most significant genes using lambda min to develop the prognostic risk model. The optimized model was: risk score = 0.104 * SLC2A1 + 0.138 * FSTL3 + 0.069 *C1QTNF6 + 0.046 * PTX3. We calculated the risk scores of each sample using this formula and classified all the samples into high- and low-risk groups according to the median for further analysis.

To validate the performance of our model, we plotted the Kaplan–Meier survival curves of the high- and low-risk groups (**Figure 3A**). A significant association was shown between the risk group and survival ($p < 0.0001$), suggesting that the model had a

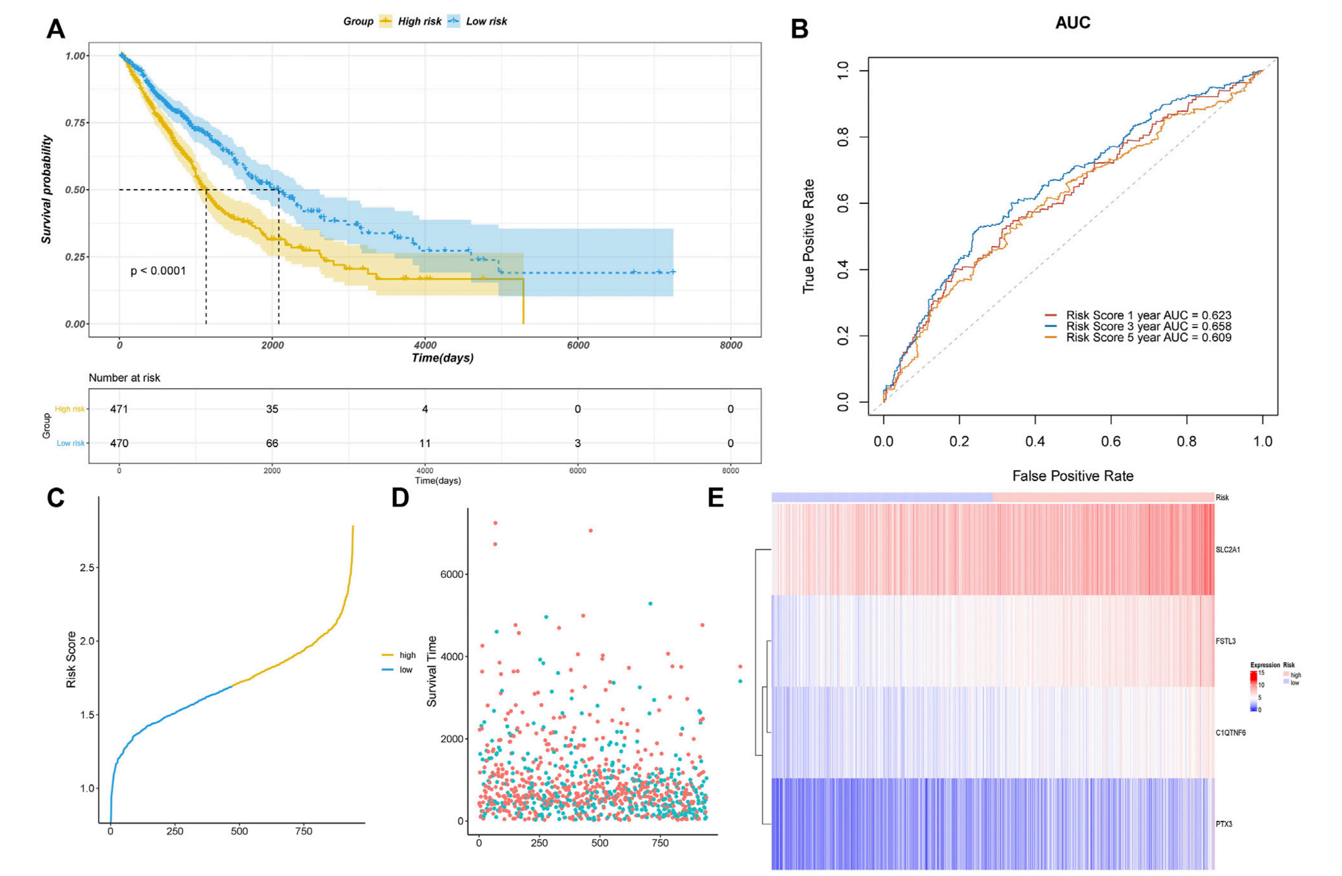


FIGURE 4 | Assessment of the risk model based on TCGA data. **(A)** Kaplan–Meier curve validation. **(B)** ROC curve validation. **(C)** Risk score of all samples. **(D)** Scatter plot of the survival time of all samples. **(E)** Heatmap of the prognostic genes in high- and low-risk groups.

high prognostic value. Time-dependent ROC curves were further plotted, which showed AUC>0.6 in the 1-year, 3-years, and 5-years ROC curves. This indicated that the model had good prediction ability (**Figure 3B**). Based on the optimistic cutoff, the patients were divided into high AF risk score and low AF risk score groups (**Figures 3C–F**).

We used the GSE4573 and GSE68465 datasets to validate the model (**Figures 4A–D**). We combined the two datasets and removed the batch effect. We selected the prognostic genes in the datasets (C1QTNF6 was not identified) and calculated the risk score using the coefficients in the model for validation. The Kaplan–Meier plot showed that the samples in the high-risk group had a worse prognosis with a p-value < 0.05, which indicated that our model had high accuracy.

2.4 Differential Analysis and Association Analysis of the Angiogenic Factor Risk Score

We analyzed the difference in AF risk scores of each group that was stratified by clinical characteristics. The risk score of LUSC was significantly higher than that of LUAD (**Figure 5A**). The risk score of the samples with EGFR mutations was significantly lower than that of samples without EGFR mutations (**Figure 5B**). The risk score also differed significantly across the different tumor stages and TNM stages, which was consistent with the process of carcinogenesis (**Figures 5C–F**). The patients with a smoking history also had significantly higher risk scores than those who never smoked (**Figure 5G**).

We also visualized the association of the risk score with tumor mutational burden (TMB), homologous recombination deficiency (HRD), neoantigen burden, chromosomal instability (CIN), and stemness index (mRNAsi) (**Figures 6A–D**). TMB is a marker for genomic instability measured by sequencing the whole tumor genome and has been shown to correlate with immunotherapy (Gibney et al., 2016). Therefore, TMB is emerging as a predictor of immunotherapeutic responses. For all indexes, the highest correlation was obtained for TMB (**Figure 6A**). This further illustrates that the interaction of AFs can affect immunotherapy. The discovery of homologous recombination deficiency (HRD) in lung cancer is of great importance for patients who will benefit from poly ADP-ribose polymerase inhibitor (PARPi) (Weston et al., 2010). However, we did not find a correlation between HRD and AFs (**Figure 2B**). Neoantigens are another important index for predicting the clinical response to immunotherapy. The current studies of neoantigen sources mainly focus on single nucleotide variants (SNVs), such as small insertions and deletions (indels), somatic copy number variations (SCNVs), and large scale transition (LSTms). Similarly, we found no

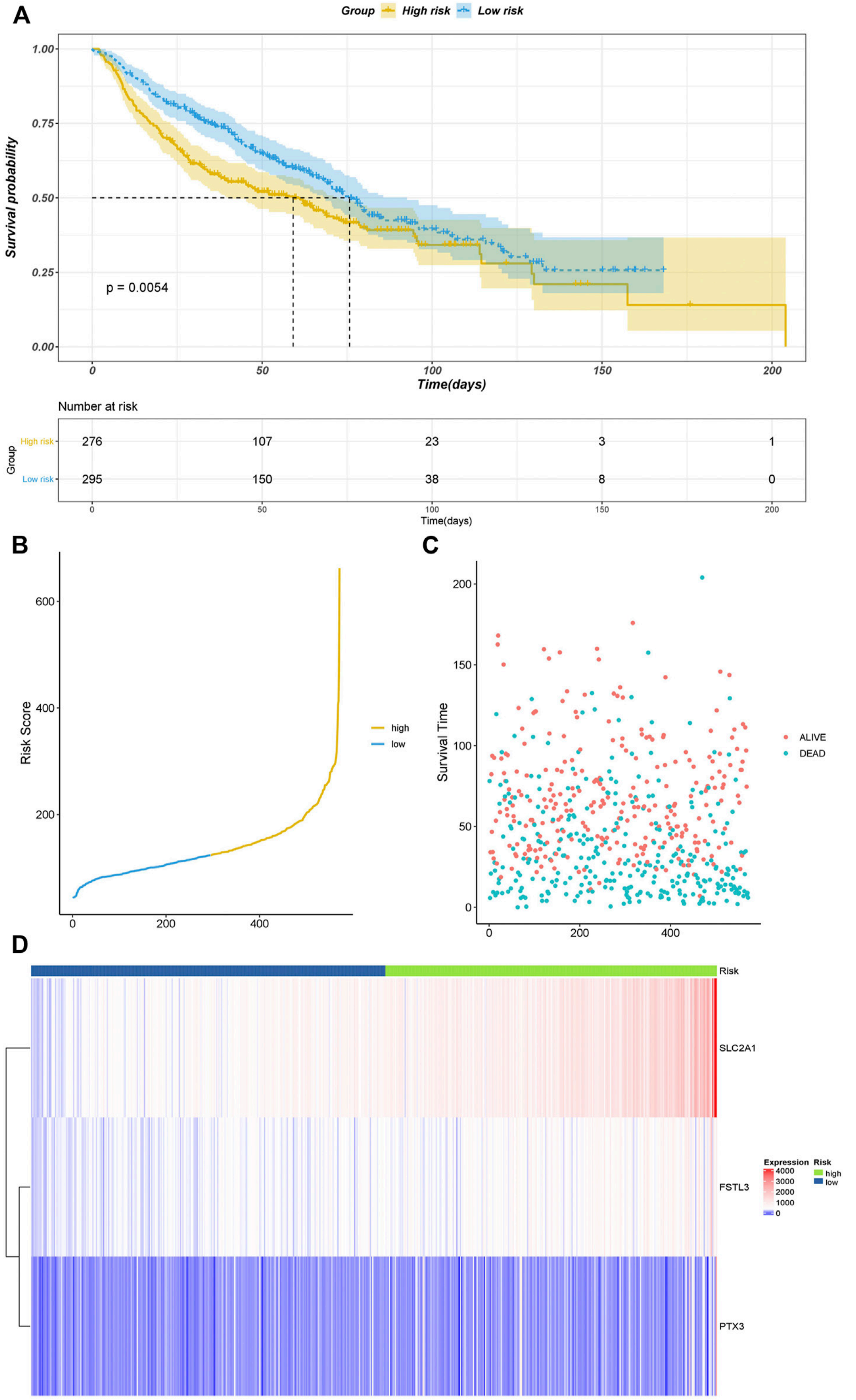


FIGURE 5 | Validation results of datasets GSE4573 and GSE68465. **(A)** Kaplan–Meier plot. **(B)** Risk score of all samples. **(C)** Scatter plot of the survival time of all samples. **(D)** Heatmap of the prognostic genes in high- and low-risk groups.

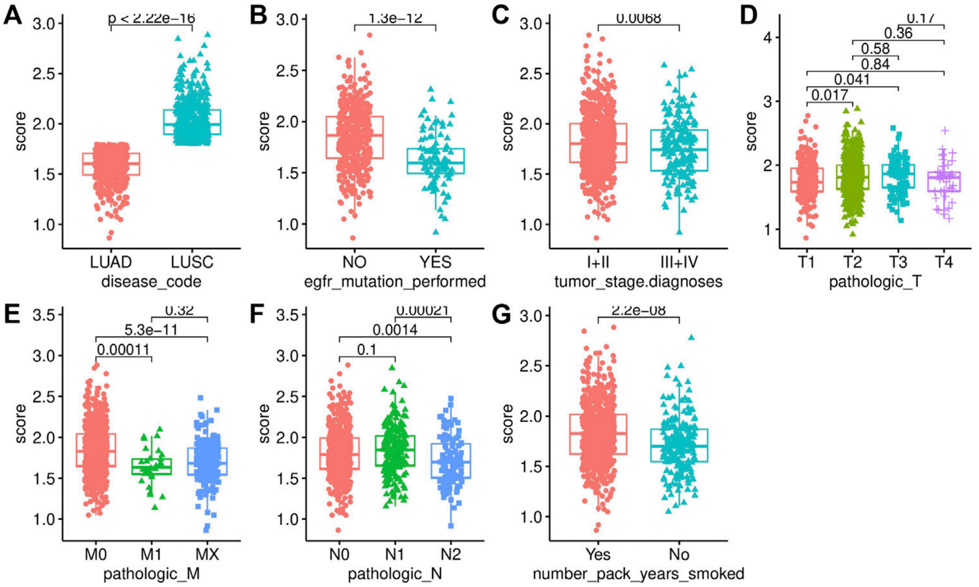


FIGURE 6 | Association analysis with clinical characteristics. **(A)** Disease code. **(B)** EGFR mutation status. **(C)** Tumor stage. **(D)** T stage. **(E)** M stage. **(F)** N stage. **(G)** Smoking history.

significant differences in these parameters (**Figures 6C–H**). The stemness index (mRNAsi) is used to measure the tumor development and evaluate the reliability of stem cell indexes as shown in **Figures 6A,I** significant positive correlation was found between AFs and mRNAsi. These results confirmed that AFs were related to biological processes, cancer metastasis, and the immune microenvironment.

2.5 Immune Infiltration Analysis of High- and Low-Risk Groups

The immune infiltration status was highly associated with the prognosis of NSCLC. We used the CIBERSORT algorithm to calculate and compare the proportion of immune infiltration in the high- and low-risk groups based on TCGA data (**Figure 7A**). The proportions of naive B cells, memory activated CD4 T cells, gammadelta T cells, and resting dendritic cells were significantly increased in the low-risk group, while the proportions of memory B cells, and macrophages. M0, macrophages. M2, and activated mast cells was significantly higher in the high-risk group, which indicated that the immune infiltration status was different in the high- and low-risk groups.

We also found that the stroma score (p = 7.8e-16), immune score (p = 0.012), and tumor purity (p = 1.7e-08) were significantly higher in the high-risk group than in the low-risk group (**Figures 7B–D**).

2.6 Differences in the Mutation Profile Between High- and Low-Risk Groups

We further investigated the difference in the mutation profiles between the high- and low-risk groups based on TCGA data. The mutation rate of the high-risk group was slightly higher than that of the low-risk group (92.81 vs. 90.91%). The mutation rate of TP53 was the highest in both the high- and low-risk groups. Additionally, missense mutations were the most dominant among all mutation types. Single nucleotide polymorphisms (SNPs) occurred more frequently in the high-risk group than in the low-risk group (**Figures 8A,B**).

We also investigated the difference in CNV between the high- and low-risk groups (**Figures 8C–E**). The copy numbers of amplification and deletion were distributed differently in the same position. Significant differences in distribution could be observed in the figures (**Figures 8C,D**). We analyzed the Z-score of the high- and low-risk groups (**Figure 8E**) by t-test. The results showed a significant difference between them (p < 2.22e-16).

2.7 Independent Prognosticator Analysis of Risk Score

Immunotherapy offers a new approach to cancer treatment. For a long period of time, immunotherapy approaches targeting PD1, PDL1, and ctla-4 have all been successfully applied in cancer, with largely promising outcomes. Tumor immune dysfunction and exclusion (TIDE) is a gene expression biomarker developed for predicting the clinical response to immune checkpoint blockade. We used the TIDE score to assess the performance of the risk score to predict the response to immunotherapy and visualized it in R software. A significant difference in the TIDE score was demonstrated between the high- and low-risk groups (p = 0.0027) (**Figure 9A**), while its prediction accuracy was lower than that of the risk score (**Figure 9B**).

To assess the effect of the risk score on prognosis, we performed univariate and multivariate Cox regression analyses

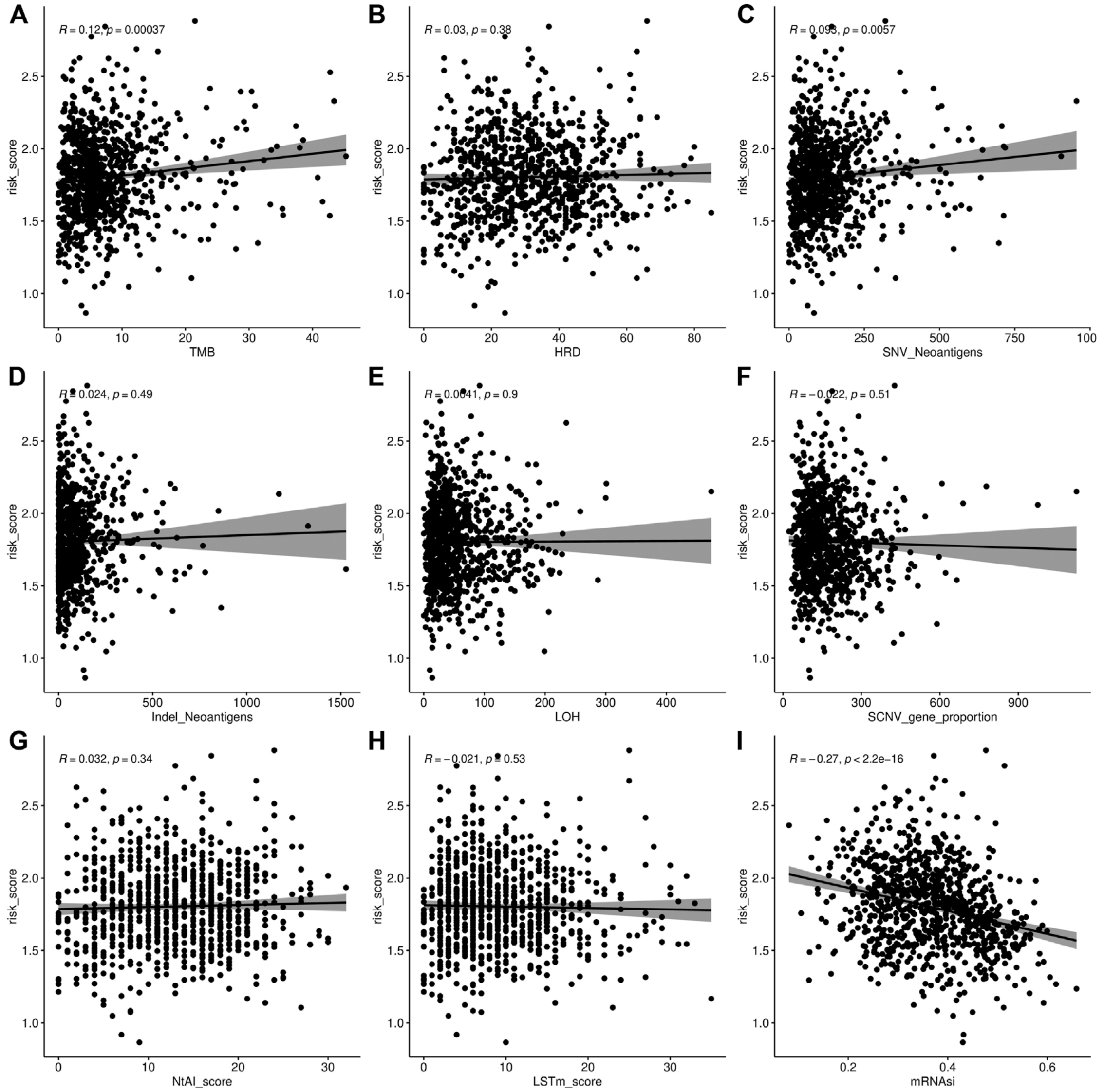


FIGURE 7 | Association analysis of AF risk score. **(A)** Tumor mutational burden and AF risk score. **(B)** Homologous recombination deficiency and AF risk score. **(C–D)** Neoantigen burden and AF risk score. **(E)** Loss of heterozygosity (LOH) in chromosome instability and AF risk score. **(F)** SCNV of chromosome instability and AF risk score. **(G)** Telomeric allelic imbalance (NtAI) of chromosome instability and AF risk score. **(H)** Large scale transition (LSTm) **(I)** Stemness index and AF risk score.

of the above clinical characteristics and validated the risk model using validation datasets (**Figures 10A–D**). The risk score showed a significant effect on the prognosis in both univariate and multivariate regression analyses.

2.8 Prognostic Analysis of Risk Score and Clinical Characteristics

Finally, we developed nomograms using the risk score and clinical characteristics and validated them with calibration plots (**Figure 11A**). The risk score showed the highest accuracy of prediction (**Figure 11B**). The 1-year, 2-years, and 3-years calibration plots demonstrated the highest accuracy of our nomograms (**Figures 11C–E**).

3 DISCUSSION

Angiogenesis is essential for tumor growth and metastasis and can provide space and nutrients for tumor cells. Multiple angiogenic growth factors play critical roles in this process. The previous studies indicate that targeting tumor angiogenesis is a promising way to fight tumor growth and dissemination in numerous types of cancer (DeBusk et al., 2010; Meng et al., 2017; Chu et al., 2021; Pan et al., 2021).

With the development of next-generation sequencing, more extensive molecules have been discovered as therapeutic targets. However, no study has previously constructed a prediction model of NSCLC based on AFs. In this study, we first identified 372 DE-AFs based on the UCSC database and then confirmed that four

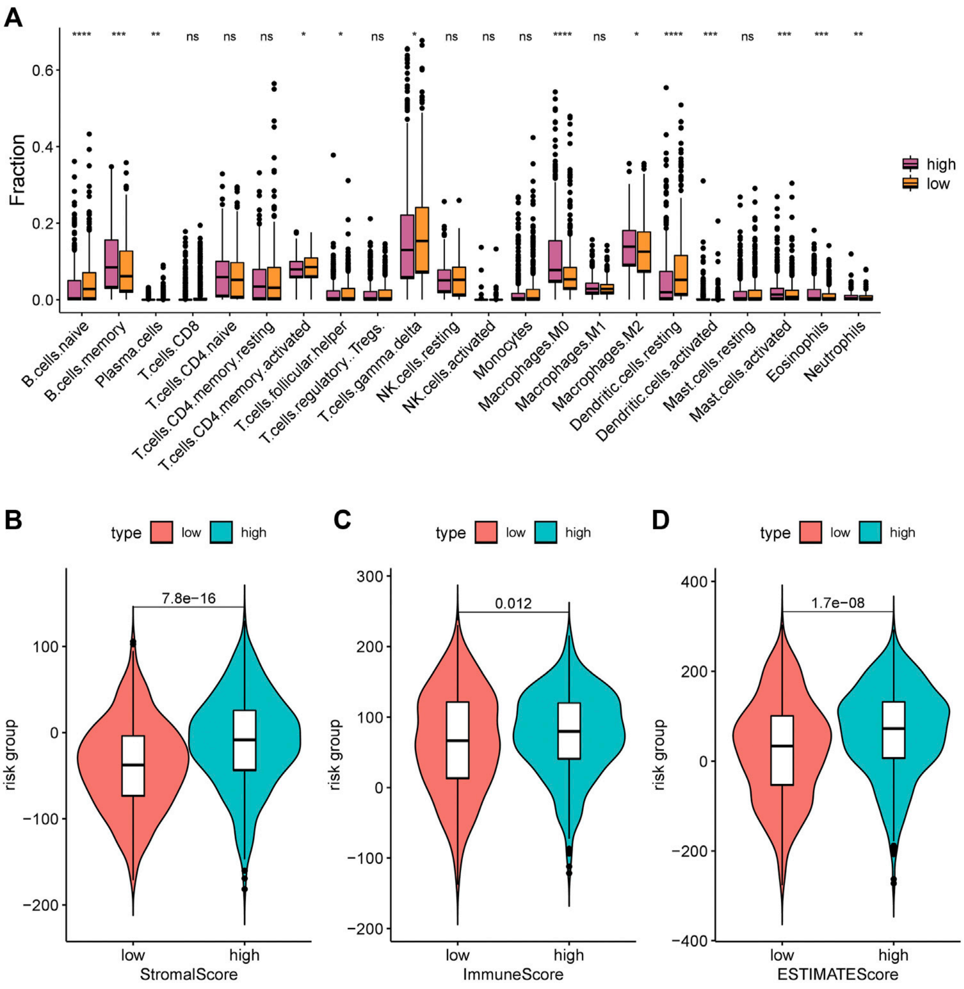


FIGURE 8 | Immune infiltration levels of 22 immune cell types in the low-risk group and high-risk group. **(A)** CIBERSORT algorithm was used to assess the difference in immune infiltration: *, *p* < 0.05; **, *p* < 0.001; ***, *p* < 0.01; ****, *p* < 0.001; ns, *p* > 0.05 (nonsignificant). **(B)** Stromal score; **(C)** Immune score; and **(D)** ESTIMATE score.

genes, C1QTNF6, SLC2A1, PTX3, and FSTL3, were significantly correlated with prognosis by constructing Cox regression and Lasso regression models. High expression of the four genes was also associated with poor prognosis in NSCLC patients. Second, according to the medium-risk score, NSCLC patients were divided into high- and low-risk groups. We calculated each AUC value of the ROC curves for predicting prognosis, which all had significantly good sensitivity. The 1-, 3-, and 5-years AUC values of the ROC were 0.623, 0.658, and 0.609, respectively. The risk score also performed well in validation sets GSE4573 and GSE68465. We also evaluated our AF risk score models on GSE4573 and GSE68465 validation data (batch effect correction). The results showed significant differences between the high- and low-risk groups.

The results of our study were consistent with those of other past studies. Wei et al. (Zhang and Feng 2021) reported that C1QTNF6 was significantly highly expressed in NSCLC tissues and cells and regulated tumor growth, migration, and apoptosis. Similar results have been reported in Japan (Tamotsu et al.) (Takeuchi et al., 2011), in which C1QTNF6 has been implicated in tumor angiogenesis in hepatocellular carcinoma. Solute carrier family 2 member 1 protein (SLC2A1) plays an important role in glucose metabolism in the human body. A previous study suggested that the upregulated expression level of SLC2A1 may increase the tumor cell proliferation and metastasis (Xiong et al., 2020). Hongwei et al. (Xia et al., 2021) found that lncRNA PVT1 can regulate cell growth, migration, and invasion by targeting the miR-378c/SLC2A1 axis. PTX3 is involved in tumor progression in multiple types of cancer and has also been identified as an independent prognostic predictor of cancer (Bonavita et al., 2015; Giacomini et al., 2018). Follistatin-related gene 3 (FSTL3) was proven to be an oncogene, and upregulated the expression of FSTL3 could activate migration by promoting F-actin and BMP/SMAD signaling (Chu et al., 2020; Liu et al., 2021). Although C1QTNF6, SLC2A1, PTX3, and FSTL3 may serve as potential targets for antiangiogenic therapeutic strategies, the molecular mechanisms of angiogenesis remain unclear.

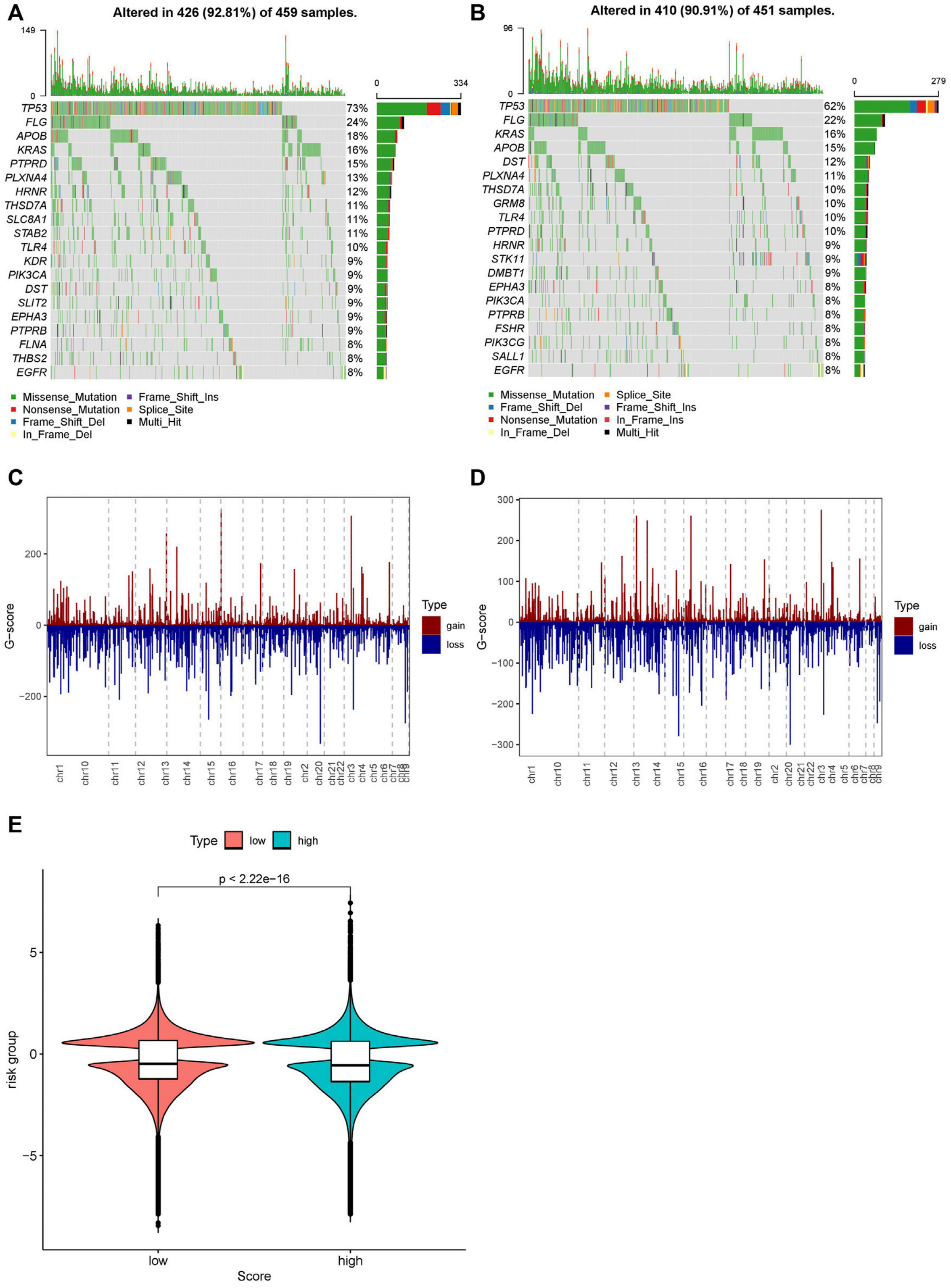


FIGURE 9 | Distribution of mutations and CNVs in the high- and low-risk groups. **(A)** Mutations in the high-risk group. **(B)** Mutations in the low-risk group. **(C)** CNVs in the low-risk group. **(D)** CNVs in the high-risk group. **(E)** Distribution of the G-score and the *p*-value of the Wilcoxon test in the high- and low-risk groups.

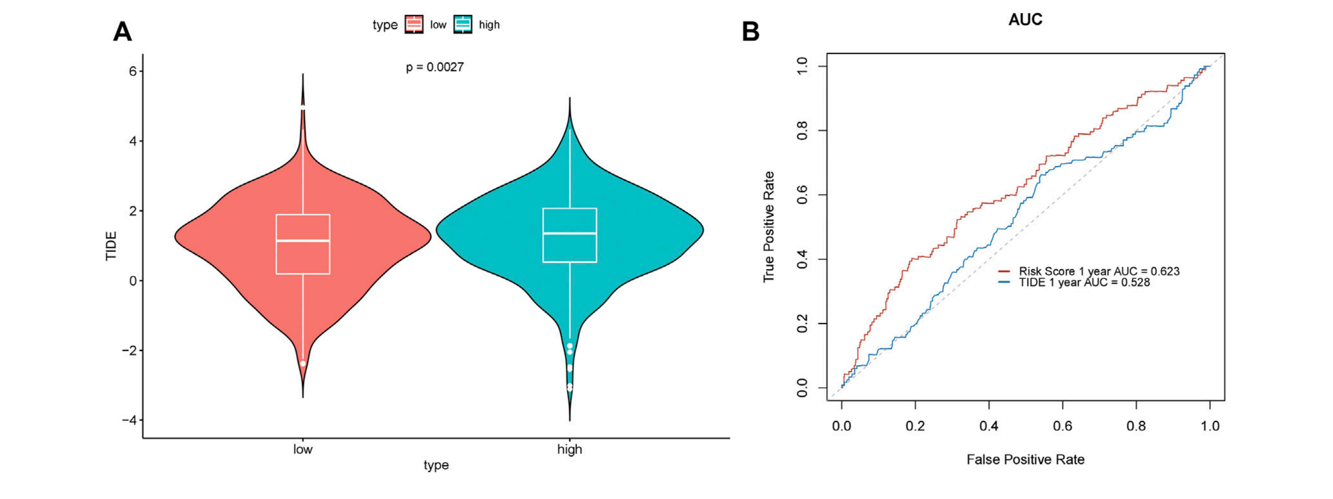


FIGURE 10 | Prediction performance of the TIDE score. **(A)** Difference in TIDE scores in the high- and low-risk groups. **(B)** ROC curve.

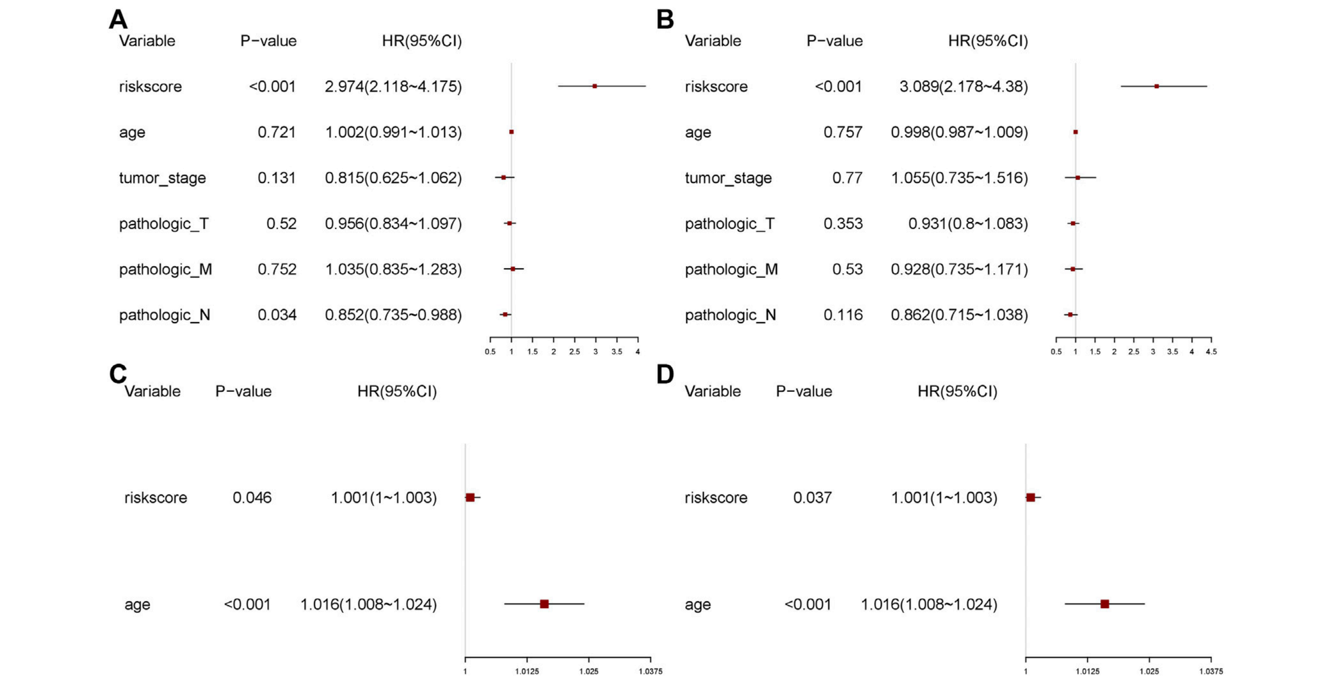


FIGURE 11 | Univariate and multivariate regression analyses. **(A)** Univariate analysis. **(B)** Multivariate analysis. **(C)** Univariate analysis of the validation set. **(D)** Multivariate analysis of the validation set.

Third, to better guide clinical decision-making, we applied AFRS to different clinical samples. We were pleasantly surprised that AFRS in LUSC patients was significantly higher than that in LUAD patients. AFRS was significantly lower in the patients with EGFR mutation or without smoking. Furthermore, we conducted a correlation between AFRS and different clinical stages and found that AFRS was closely related to the clinical stage and TNM stage.

Fourth, in recent research, immunotherapy has been increasingly recognized for its potential therapeutic effect on a variety of tumors. For example, immune checkpoint (PD-1, CTLA-4) blockade has become an increasingly important part of cancer therapy (Passiglia et al., 2021). There were plenty of clinical trials (Reck et al., 2019; Herbst et al., 2020; Patel et al., 2020) that proved the combination of ICI therapy and angiogenesis therapy can reprogram the immune microenvironment and prune cancer

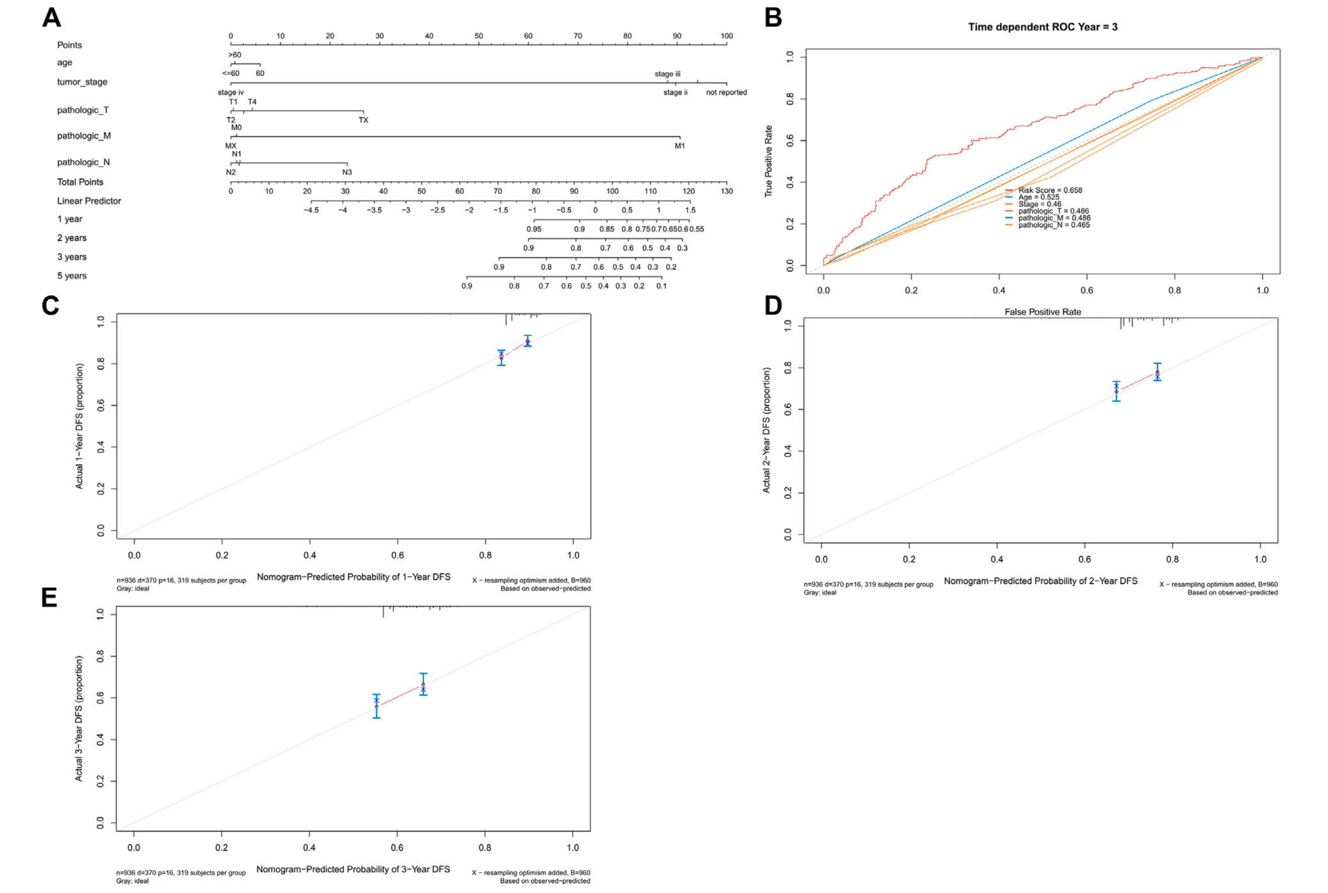


FIGURE 12 | Nomogram and calibration plots. **(A)** Nomogram of age, tumor stage, and TNM stage. **(B)** ROC curve of risk score, age, tumor stage, and TNM stage. **(C)** 1-year calibration plot. **(D)** 2-years calibration plot and **(E)** 3-years calibration plot.

growth-related blood vessels (Ramjiawan et al., 2017; Yi et al., 2019; Giannone et al., 2020), which could have a synergistically better performance in prolonging overall survival, especially in patients with activating mutations of EGFR (Reck et al., 2019). By detecting the immunity indexes of TMB and mRNAsi, we believed that this research might provide bioinformatics evidence to support the design of a combination of immunotherapy and antiangiogenic therapy for NSCLC patients in the future.

Fifth, we found that of all clinical samples, the TP53 mutation type had the highest rate of mutations, neither in the low nor high AFRS group. The SNP mutation in the high AFRS group was remarkably higher than that in the low AFRS group. Numerous studies have proven that TP53 mutation in cancers can influence drug activity, tumor apoptosis, and immune evasion (Alexandrova et al., 2015; Srihari et al., 2018). Notably, gain-of-function p53 mutation promotes neutrophils to tumors, which leads to resistance to immunotherapy (Siolas et al., 2021). As a result, we further analyzed the correlation of AFRS with the infiltration of various immune cells. We found that the immune response was significantly altered between the low and high AFRS groups, including immune cell infiltration (i.e., M2 macrophages and M0, mast cells, B cells), immune score, stromal score, and ESTIMATE score. These results indicated that the high AFRS group could induce stronger immunity activity.

For better clinical applications, we strive to develop a nomogram to predict the prognosis of NSCLC patients. The established nomogram showed a great performance in predicting the clinical outcomes for NSCLC patients.

However, this study has several limitations. First, due to limited resources and funding available, no clinical samples were analyzed, hence, clinical relevance was not assessed. Second, the lack of experimental verification was also limited. We will further confirm our conclusions by performing cell line and animal model experiments in the future and prove the changes in the protein levels by western blot analysis.

4 CONCLUSION

In conclusion, assessing the global gene expression profile of Afs in this study was the first. From the perspective of a reliable risk score model using angiogenic factors, the present study provided a new method for NSCLC treatment in the clinic. However, the established model needs to be further confirmed in the future by large scale multicenter clinical studies.

5 MATERIALS AND METHODS

5.1 Sources of Non-Small-Cell Lung Cancer Datasets

The expression profile combined with patients' clinical and annotation information in LUAD and LUSC datasets were downloaded from UCSC (https://xenabrowser.net/). Next, we averaged the expression level of genes with the same name and removed the genes with expression levels lower than 30%. We merged the two expression profiles after processing and converted the data type from FPKM to TPM. The samples from patients aged >18 years were extracted, and batch effects were removed (**Figures 12A,B**). We then searched the NCBI database (https://www.ncbi.nlm.nih.gov/gene/?term=angiogenic) using "angiogenic factor (AF)" as the keyword and extracted AF expression data of 1,054 samples from the downloaded expression profile.

5.2 Enrichment Analysis of Angiogenic Factors Expression

We used the R package "limma" to identify AF-related differentially expressed genes (DEGs) (threshold: adj.P.Val<0.01 & |log (FC) |≥ 1) in 372 cancer and normal samples. Next, gene ontology (GO) enrichment analysis (*p*-value cutoff < 0.05) and KEGG pathway enrichment analysis (*p*-value cutoff < 0.05) of differentially expressed genes were performed using the R package "clusterProfiler".

5.3 Univariate Cox Regression Analysis

Other data of cancer samples were further extracted, and a univariate Cox regression analysis of DEGs associated with overall survival was performed using the R packages "survival" and "survminer" with a threshold of $p < 0.05$. DEGs associated with prognosis were identified after screening.

5.4 Prognostic Risk Model Development Based on Lasso Regression

Lasso regression was performed using the R package "glmnet" for downscaling prognostic genes. We first screened lambda by cross-validation, and then selected the model with lambda. min. Next, the expression matrix of the selected genes for the model was extracted, and the risk score of each sample was calculated using the following formula:

$$\mathrm{Riskscore_i} = \sum_{i=1}^{n} \mathrm{exp_{ji}} * \beta_j.$$

It represented the expression level of gene j in sample i, and represented the coefficient of gene j in the lasso regression model. All the samples were stratified into high- and low-risk groups according to the median-risk score.

5.5 Risk Model Assessment

Kaplan–Meier survival curves were plotted according to high- or low-risk groups. The ROC curves were drawn based on the predicted risk score of each sample.

5.6 Analysis of Angiogenic Factor Risk Scores According to Clinical Characteristics

The samples with AF risk scores were grouped according to clinical characteristics. We used the R package "ggplot2" to show the distribution of AF risk scores in each group and the R package "ggpubr" to illustrate the significant difference between groups.

5.7 Association Analysis of Angiogenic Factor Risk Score

We calculated tumor mutational burden, homologous recombination deficiency (HRD) (from technical support), tumor neoantigen burden (according to the literature The Immune Landscape of Cancer), chromosome instability (CIN) (according to the literature The Immune Landscape of Cancer), and stemness index (according to the literature Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation) based on AF risk scores and performed association analyses.

5.8 Assessment of Immune Infiltration in the High- and Low-Risk Groups Using CIBERSORT

The proportion of 22 immune cells in the samples can be derived using the CIBERSORT algorithm based on the expression of certain genes. We assessed the difference in immune infiltration between the high- and low-risk groups by *t*-test with a significance threshold of $p < 0.05$.

5.9 Assessment of Immune Score, Stromal Score, and Tumor Purity Using ESTIMATE

We analyzed the differences in the immune score, stromal score, and tumor purity of AF in high- and low-risk groups using the R package "ESTIMATE" and assessed the differences in immune infiltration in the high- and low-risk groups by *t*-test with a threshold of $p < 0.05$.

5.10 Mutation Analysis in High and Low-Risk Groups

MAF files of NSCLC were downloaded from the GDC database, and we extracted the mutation information of AF from the somatic mutation profile. The mutation profile of AF in high- and low-risk groups was demonstrated with the help of the "oncoplot" function, using the R package "maftools".

5.11 Copy Number Variation Analysis of High- and Low-Risk Groups

The copy number variation (CNV) data of LUSC and LUAD were downloaded from UCSC. The copy numbers of the high- and low-risk groups were extracted to generate files and plotted on the gadget using the "CNV distribution chart - bar graph section".

5.12 Prediction of Response to Immunotherapy

The expression profiles of immune genes were extracted from the processed data of endometrial cancer samples, and the immune gene sets were obtained from the ImmPort database (https://www.immport.org/) and InnateDB database (https://www.innatedb.ca/). The expression profiles of the immune gene sets were subsequently normalized. The predicted TIDE scores of the samples were calculated using the TIDE online database. The distribution of TIDE scores in the high- and low-risk groups was illustrated with box plots using ggpubr, and the significance was tested by *t*-test.

5.13 Independent Prognostic Factor Analysis

To validate whether the risk score was an independent prognostic factor, univariate Cox regression analyses of the candidate prognostic factors using TCGA sample data were first performed, including risk score, age, tumor stage, and TNM stage. A multivariate Cox regression analysis was subsequently performed to assess the effect size of the risk score. We used the function cph in the R package "rms" to plot the nomograms and calibration plots for visualization.

5.14 Statistical Analysis

All statistical analyses were performed using R software version 4.0.3. $p < 0.05$ was set as the significance criterion.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

XG and YK designed the concept of the study. Bioinformatics analysis and statistical analysis were performed by YK. XG and LC wrote the draft of the manuscript. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.894024/full#supplementary-material

REFERENCES

Alexandrova, E. M., Yallowitz, A. R., Li, D., Xu, S., Schulz, R., Proia, D. A., et al. (2015). Improving Survival by Exploiting Tumour Dependence on Stabilized Mutant P53 for Treatment. *Nature* 523 (7560), 352–356. doi:10.1038/nature14430

Bonavita, E., Gentile, S., Rubino, M., Maina, V., Papait, R., Kunderfranco, P., et al. (2015). PTX3 Is an Extrinsic Oncosuppressor Regulating Complement-dependent Inflammation in Cancer. *Cell* 160 (4), 700–714. doi:10.1016/j.cell.2015.01.004

Chatterjee, N., and Bivona, T. G. (2019). Polytherapy and Targeted Cancer Drug Resistance. *Trends Cancer* 5 (3), 170–182. doi:10.1016/j.trecan.2019.02.003

Chu, L., Li, N., Deng, J., Wu, Y., Yang, H., Wang, W., et al. (2020). LC-APCI+-MS/MS Method for the Analysis of Ten Hormones and Two Endocannabinoids in Plasma and Hair from the Mice with Different Gut Microbiota. *J. Pharm. Biomed. Anal.* 185, 113223. doi:10.1016/j.jpba.2020.113223

Chu, L., Huang, Y., Xu, Y., Wang, L.-K., and Lu, Q. (2021). An LC-APCI+-MS/MS-based Method for Determining the Concentration of Neurosteroids in the Brain of Male Mice with Different Gut Microbiota. *J. Neurosci. Methods* 360, 109268. doi:10.1016/j.jneumeth.2021.109268

DeBusk, L. M., Boelte, K., Min, Y., and Lin, P. C. (2010). Heterozygous Deficiency of δ-catenin Impairs Pathological Angiogenesis. *J. Exp. Med.* 207 (1), 77–84. doi:10.1084/jem.20091097

Ettinger, D. S., Akerley, W., Borghaei, H., Chang, A. C., Cheney, R. T., Chirieac, L. R., et al. (2013). Non-Small Cell Lung Cancer, Version 2.2013. *J. Natl. Compr. Canc Netw.* 11 (6), 645–653. doi:10.6004/jnccn.2013.0084

Ettinger, D. S., Wood, D. E., Aisner, D. L., Akerley, W., Bauman, J. R., Bharat, A., et al. (2021). NCCN Guidelines Insights: Non-small Cell Lung Cancer, Version 2.2021. *J. Natl. Compr. Cancer Netw.* 19 (3), 254–266. doi:10.6004/jnccn.2021.0013

Ferrara, N., Hillan, K. J., and Novotny, W. (2005). Bevacizumab (Avastin), a Humanized Anti-VEGF Monoclonal Antibody for Cancer Therapy. *Biochem. Biophysical Res. Commun.* 333 (2), 328–335. doi:10.1016/j.bbrc.2005.05.132

Giacomini, A., Ghedini, G. C., Presta, M., and Ronca, R. (2018). Long Pentraxin 3: A Novel Multifaceted Player in Cancer. *Biochim. Biophys. Acta (Bba) - Rev. Cancer* 1869 (1), 53–63. doi:10.1016/j.bbcan.2017.11.004

Giannone, G., Ghisoni, E., Genta, S., Scotto, G., Tuninetti, V., Turinetto, M., et al. (2020). Immuno-Metabolism and Microenvironment in Cancer: Key Players for Immunotherapy. *Ijms* 21 (12), 4414. doi:10.3390/ijms21124414

Gibney, G. T., Weiner, L. M., and Atkins, M. B. (2016). Predictive Biomarkers for Checkpoint Inhibitor-Based Immunotherapy. *Lancet Oncol.* 17 (12), e542–e551. doi:10.1016/S1470-2045(16)30406-5

Goveia, J., Rohlenova, K., Taverna, F., Treps, L., Conradi, L.-C., Pircher, A., et al. (2020). An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell* 37 (1), 21–36. e13. doi:10.1016/j.ccell.2019.12.001

Gridelli, C., Rossi, A., Carbone, D. P., Guarize, J., Karachaliou, N., Mok, T., et al. (2015). Non-small-cell Lung Cancer. *Nat. Rev. Dis. Primers* 1 (1), 1. doi:10.1038/nrdp.2015.9

Hanahan, D., and Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell* 100 (1), 57–70. doi:10.1016/s0092-8674(00)81683-9

Herbst, R. S., Giaccone, G., de Marinis, F., Reinmuth, N., Vergnenegre, A., Barrios, C. H., et al. (2020). Atezolizumab for First-Line Treatment of PD-L1-Selected Patients with NSCLC. *N. Engl. J. Med.* 383 (14), 1328–1339. doi:10.1056/NEJMoa1917346

Hirsch, F. R., Scagliotti, G. V., Mulshine, J. L., Kwon, R., Curran, W. J., Jr., Wu, Y.-L., et al. (2017). Lung Cancer: Current Therapies and New Targeted Treatments. *The Lancet* 389 (10066), 299–311. doi:10.1016/S0140-6736(16)30958-8

Liu, Y.-J., Li, J.-P., Zhang, Y., Nie, M.-J., Zhang, Y.-H., Liu, S.-L., et al. (2021). FSTL3 Is a Prognostic Biomarker in Gastric Cancer and Is Correlated with M2 Macrophage Infiltration. *Ott* 14, 4099–4117. doi:10.2147/ott.S314561

Lugano, R., Ramachandran, M., and Dimberg, A. (2020). Tumor Angiogenesis: Causes, Consequences, Challenges and Opportunities. *Cell. Mol. Life Sci.* 77 (9), 1745–1770. doi:10.1007/s00018-019-03351-7

Ma, J., Wang, J., Ghoraie, L. S., Men, X., Haibe-Kains, B., and Dai, P. (2019). Network-based Approach to Identify Principal Isoforms Among Four Cancer Types. *Mol. Omics* 15 (2), 117–129. doi:10.1039/c8mo00234g

Markham, A., and Keam, S. J. (2019). Camrelizumab: First Global Approval. *Drugs* 79 (12), 1355–1361. doi:10.1007/s40265-019-01167-0

Meng, J., Liu, Y., Han, J., Tan, Q., Chen, S., Qiao, K., et al. (2017). Hsp90β Promoted Endothelial Cell-dependent Tumor Angiogenesis in Hepatocellular Carcinoma. *Mol. Cancer* 16 (1), 72. doi:10.1186/s12943-017-0640-9

Oxnard, G. R., Arcila, M. E., Chmielecki, J., Ladanyi, M., Miller, V. A., and Pao, W. (2011). New Strategies in Overcoming Acquired Resistance to Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors in Lung Cancer. *Clin. Cancer Res.* 17 (17), 5530–5537. doi:10.1158/1078-0432.Ccr-10-2571

Pan, S., Zhao, X., Shao, C., Fu, B., Huang, Y., Zhang, N., et al. (2021). STIM1 Promotes Angiogenesis by Reducing Exosomal miR-145 in Breast Cancer MDA-MB-231 Cells. *Cell Death Dis* 12 (1), 38. doi:10.1038/s41419-020-03304-0

Passiglia, F., Galvano, A., Gristina, V., Barraco, N., Castiglia, M., Perez, A., et al. (20212021). Is There Any Place for PD-1/CTLA-4 Inhibitors Combination in the First-Line Treatment of Advanced NSCLC?-a Trial-Level Meta-Analysis in PD-L1 Selected Subgroups. *Transl Lung Cancer Res.* 10 (7), 3106–3119. doi:10.21037/tlcr-21-52

Patel, J. D., Lee, J.-W., Carbone, D. P., Wagner, H., Shanker, A., de Aquino, M. T. P., et al. (2020). Phase II Study of Immunotherapy with Tecemotide and Bevacizumab after Chemoradiation in Patients with Unresectable Stage III Non-squamous Non-small-cell Lung Cancer (NS-NSCLC): A Trial of the ECOG-ACRIN Cancer Research Group (E6508). *Clin. Lung Cancer* 21 (6), 520–526. doi:10.1016/j.cllc.2020.06.007

Ramjiawan, R. R., Griffioen, A. W., and Duda, D. G. (2017). Anti-angiogenesis for Cancer Revisited: Is There a Role for Combinations with Immunotherapy? *Angiogenesis* 20 (2), 185–204. doi:10.1007/s10456-017-9552-y

Reck, M., Mok, T. S. K., Nishio, M., Jotte, R. M., Cappuzzo, F., Orlandi, F., et al. (2019). Atezolizumab Plus Bevacizumab and Chemotherapy in Non-small-cell Lung Cancer (IMpower150): Key Subgroup Analyses of Patients with EGFR Mutations or Baseline Liver Metastases in a Randomised, Open-Label Phase 3 Trial. *Lancet Respir. Med.* 7 (5), 387–401. doi:10.1016/S2213-2600(19)30084-0

Siolas, D., Vucic, E., Kurz, E., Hajdu, C., and Bar-Sagi, D. (2021). Gain-of-function p53R172H Mutation Drives Accumulation of Neutrophils in Pancreatic Tumors, Promoting Resistance to Immunotherapy. *Cel Rep.* 36 (8), 109578. doi:10.1016/j.celrep.2021.109578

Srihari, S., Kwong, R., Tran, K., Simpson, R., Tattam, P., and Smith, E. (2018). Metabolic Deregulation in Prostate Cancer. *Mol. Omics* 14 (5), 320–329. doi:10.1039/c8mo00170g

Takeuchi, T., Adachi, Y., and Nagayama, T. (2011). Expression of a Secretory Protein C1qTNF6, a C1qTNF Family Member, in Hepatocellular Carcinoma. *Anal. Cell Pathol.* 34 (3), 113–121. doi:10.3233/ACP-2011-00910.1155/2011/578097

Wang, M., An, S., Wang, D., Ji, H., Geng, M., Guo, X., et al. (2019). Quantitative Proteomics Identify the Possible Tumor Suppressive Role of Protease-Activated Receptor-4 in Esophageal Squamous Cell Carcinoma Cells. *Pathol. Oncol. Res.* 25 (3), 937–943. doi:10.1007/s12253-018-0395-7

Weston, V. J., Oldreive, C. E., Skowronska, A., Oscier, D. G., Pratt, G., Dyer, M. J. S., et al. (2010). The PARP Inhibitor Olaparib Induces Significant Killing of ATM-Deficient Lymphoid Tumor Cells *In Vitro* and *In Vivo*. *Blood* 116 (22), 4578–4587. doi:10.1182/blood-2010-01-265769

Xia, H., Zhang, Z., Yuan, J., and Niu, Q. (2021). The lncRNA PVT1 Promotes Invasive Growth of Lung Adenocarcinoma Cells by Targeting miR-378c to Regulate SLC2A1 Expression. *Hum. Cel* 34 (1), 201–210. doi:10.1007/s13577-020-00434-7

Xiong, Y., Lei, J., Zhao, J., Lu, Q., Feng, Y., Qiao, T., et al. (2020). A Gene-Based Survival Score for Lung Adenocarcinoma by Multiple Transcriptional Datasets Analysis. *BMC Cancer* 20 (1), 1046. doi:10.1186/s12885-020-07473-1

Yi, M., Jiao, D., Qin, S., Chu, Q., Wu, K., and Li, A. (2019). Synergistic Effect of Immune Checkpoint Blockade and Anti-angiogenesis in Cancer Treatment. *Mol. Cancer* 18 (1), 60. doi:10.1186/s12943-019-0974-6

Zhang, H. (2015). Apatinib for Molecular Targeted Therapy in Tumor. *Dddt* 9, 6075–6081. doi:10.2147/Dddt.S97235

Zhang, W., and Feng, G. (2021). C1QTNF6 Regulates Cell Proliferation and Apoptosis of NSCLC *In Vitro* and *In Vivo*. *Biosci. Rep.* 41 (1), BSR20201541. doi:10.1042/BSR20201541

Zhao, S., Ren, S., Jiang, T., Zhu, B., Li, X., Zhao, C., et al. (2019). Low-Dose Apatinib Optimizes Tumor Microenvironment and Potentiates Antitumor Effect of PD-1/pd-L1 Blockade in Lung Cancer. *Cancer Immunol. Res.* 7 (4), canimm.0640.2017–643. doi:10.1158/2326-6066.CIR-17-0640

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

GLOSSARY

AFRS: angiogenic factors related risk score

AFs: angiogenic factors

CIN: chromosomal instability

CNV: copy number variation

DEGs: differentially expressed genes

EGFR: epidermal growth factor receptor

HRD: homologous recombination deficiency

LASSO: least absolute shrinkage and selection operator

WHO: World Health Organization

LSTm: large scale transition

NSCLC: non-small-cell lung cancer

SCLC: small cell lung cancer

SCNV: somatic copy number variations

SNPs: single nucleotide polymorphisms

TMB: tumor mutational burden

VEGF: vascular endothelial growth factor

VEGFR2: vascular endothelial growth factor receptor 2

frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 01 July 2022
doi: 10.3389/fgene.2022.952335

High Prolyl 4-Hydroxylase Subunit Alpha 3 Expression as an Independent Prognostic Biomarker and Correlated With Immune Infiltration in Gastric Cancer

Xiaoji Niu[1,2]*, Liman Ren*[3]*, Shoumei Wang*[2]*, Dong Gao*[1]*, Mingyue Ma*[2]*, Aiyan Hu*[2]*, Hongjun Qi*[1]** and Shuhui Zhang*[2]***

[1]Department of Gastroenterology of Traditional Chinese Medicine, Qinghai Province Hospital of Traditional Chinese Medicine, Xining, China, [2]Department of Pathology, Yueyang Hospital of Integrated Traditional Chinese and Western Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Department of Endocrinology, Qinghai Province Hospital of Traditional Chinese Medicine, Xining, China

OPEN ACCESS

Edited by:
Saurav Mallik,
Harvard University, United States

Reviewed by:
Loveleen Gaur,
Amity University, India
Ruifeng Hu,
Harvard Medical School,
United States

****Correspondence:***
Hongjun Qi
docqhj@sina.com
Shuhui Zhang
shzhang@126.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *25 May 2022*
Accepted: *14 June 2022*
Published: *01 July 2022*

Citation:
Niu X, Ren L, Wang S, Gao D, Ma M, Hu A, Qi H and Zhang S (2022) High Prolyl 4-Hydroxylase Subunit Alpha 3 Expression as an Independent Prognostic Biomarker and Correlated With Immune Infiltration in Gastric Cancer.
Front. Genet. 13:952335.
doi: 10.3389/fgene.2022.952335

Background: Gastric cancer (GC) has a high mortality rate and is particularly prevalent in China. The extracellular matrix protein, prolyl 4-hydroxylase subunit alpha 3 (P4HA3), has been implicated in various cancers. We aimed to assess the diagnostic and prognostic value of P4HA3 in GC and investigate its correlation with immune cell infiltration.

Methods: The present study used microarray data from the Cancer Genome Atlas (TCGA) to analyze the association of P4HA3 expression with clinicopathological features. Data from the Gene Expression Omnibus (GEO) were used for validation. Receiver operating characteristic (ROC) and Kaplan–Meier curves were constructed to determine the diagnostic and prognostic value of P4HA3 in GC. Univariate and multivariate regression analyses were performed to assess the impact of P4HA3 on overall survival (OS) rates. A protein–protein interaction (PPI) network was generated and functional enrichment evaluated. Single-sample gene set enrichment analysis (ssGSEA) was conducted to correlate P4HA3 expression with immune cell infiltration. The correlation between P4HA3 and immune check point genes was studied.

Results: P4HA3 was over-expressed in GC, along with 15 other types of cancer, including breast invasive carcinoma and cholangiocarcinoma. P4HA3 showed high diagnostic and prognostic value in GC and was an independent prognostic factor. P4HA3, TNM (tumor, node, metastases) stage, pathological stage and age all correlated with OS rates. Genes related to P4HA3 were enriched in the lumen of the endoplasmic reticulum and included procollagen-proline 3-dioxygenase activity. P4HA3 expression correlated with numbers of macrophages, natural killer (NK) cells, immature dendritic cells (iDC), mast cells, eosinophils, effective memory T cells (Tem), T-helper 1 (Th1) cells and dendritic cells (DC). P4HA3 was positively correlated with hepatitis A virus cellular receptor 2 (HAVCR2) and programmed cell death 1 ligand 2 (PDCD1LG2).

Conclusion: P4HA3 is a potential independent biomarker for prognosis of GC and may be an immunotherapy target in the treatment of GC.

Keywords: P4HA3, gastric cancer, bioinformatics, prognosis, immune infiltration

INTRODUCTION

Data from Global Cancer Statistics (https://gco.iarc.fr/today/online-analysis) indicates that gastric cancer (GC) is the fifth most frequently diagnosed cancer and the fourth leading cause of cancer-related death worldwide. Global age-standardized incidence and mortality rates are 11.1 per 100,000 and 7.7 per 100,000. Rates are considerably higher in China, where incidence and mortality occur at 20.6 per 100,000 and 15.9 per 100,000, respectively (Machlowska et al., 2020; Sung et al., 2021; Niu et al., 2022). GC is thus a significant health and economic burden worldwide and this is particularly the case in China.

The majority of GC cases are diagnosed at the late stage, resulting in a poor prognosis. However, advances in molecular biology techniques allow us to approach an understanding of precise molecular mechanisms of carcinogenesis which holds promise for development of diagnostic, prognostic and therapeutic strategies. It is known that immune-related mechanisms and markers participate in the occurrence and development of GC and appropriately targeted therapy looks promising for its treatment (Güthle et al., 2020). Such observations highlight the urgent need to identify new immune-related biomarkers to facilitate early GC diagnosis and treatment.

Prolyl 4-hydroxylase subunit alpha 3 (P4HA3) is a catalytic subunit involved in collagen synthesis. Its overexpression has been associated with tumors and with non-cancerous diseases, including idiopathic pituitary adenoma, melanoma, stomach carcinoma, breast cancer and pulmonary fibrosis (Luo et al., 2015; Song et al., 2018; Atkinson et al., 2019; Long et al., 2019; Gu et al., 2020).

A recent study has suggested that upregulation of P4HA3 is associated with enhanced metastasis and poor survival of GC patients (Song et al., 2018). However, any correlation with immune cell infiltration has been little scrutinized. The current study investigated P4HA3 expression in GC and its relationship with immune cell infiltration.

MATERIAL AND METHODS

Microarray Datasets

The Cancer Genome Atlas (TCGA) project (https://www.cancer.gov/tcga) is an open database which aims to make molecular data characterizing the cancer-related genome freely available and to link genomic data to patients' clinicopathological information. RNA-sequencing data (level 3) with corresponding clinicopathological information were downloaded from the TCGA database. Data was converted from fragments per kilobase per million (FPKM) to transcripts per million reads (TPM). Survival data were published in Cell (Liu et al., 2018). For analyses across many tumor types, TCGA and Genotype-Tissue Expression Project (GTEx), TPM-formatted RNAseq data processed by Toil were downloaded from University of California Santa Cruz (UCSC) XENA (https://xenabrowser.net/datapages/) (Vivian et al., 2017). GSE54129 and GSE103236 datasets were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) to validate the P4HA3 expression level. **Table 1** shows the details of expression datasets involved in the study.

TABLE 1 | Details of Expression Datasets for the study.

Data source	ID	Platform	Samples (cancer vs. Normal)
TCGA	-	-	375 vs. 32
GEO	GSE54129	GPL570	111 vs. 21
GEO	GSE103236	GPL4133	10 vs. 9

Diagnostic and Prognostic Value of P4HA3

Receiver operating characteristic (ROC) and Kaplan-Meier survival curves were constructed and used to analyze the diagnostic and prognostic value of P4HA3, respectively. The association between P4HA3 expression and overall survival (OS) rates of GC patients was assessed by univariate and multivariate regression analyses.

PPI Network Construction and Functional Enrichment Analysis

Protein–protein interaction (PPI) network analysis of P4HA3 was performed by using the search tool of a single named protein with default parameters within the STRING database (version 11.5, accessed date: 02 June 2022) (Szklarczyk et al., 2021). Pathway and process enrichment analysis were conducted with the following ontology sources: Gene Ontology (GO) Biological Processes, GO Cellular Components, GO Molecular Functions, Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway, Reactome Gene Sets and Canonical Pathways within the Metascape database (https://metascape.org/gp/index.html) (Hochberg and Benjamini, 1990; Zhou et al., 2019). The complete proteome were regarded as the background to enrichment. Terms with a p-value less than 0.05, a minimum count of 3 and an enrichment factor (ratio of observed counts: counts expected by chance) of more than 1.5 were acquired and grouped into clusters based on their connections. p-values were calculated from the cumulative hypergeometric distribution and q-values using the Banjamini-Hochberg procedure to account for multiple testing. Kappa scores were used as the similarity metric when performing hierachical clustering of the enriched terms and sub-trees with a similarity of >0.3 were considered a cluster. The most statistically significant term within a cluster was used to represent that cluster.

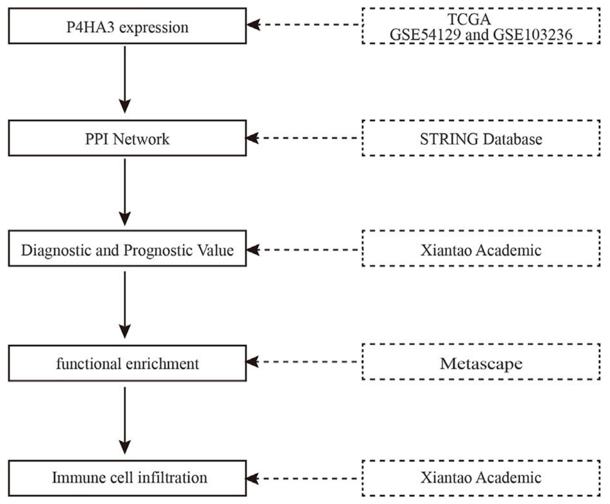


FIGURE 1 | The flowchart of the present study.

Correlation Analysis of Immune Cell Infiltration

ssGSEA was used to determine relationships between P4HA3 expression and 24 kinds of immune cells, including activated DC, B cells, macrophages and mast cells (Bindea et al., 2013). Spearman correlation analysis was used to evaluate the correlation between P4HA3 and immune cell infiltration and values of $r > 0.3$ or $r < -0.3$ and $p < 0.05$ were considered to indicate significant positive or negative correlation. The expression of the following immune-checkpoint–relevant transcripts was assessed: sialic acid binding Ig like lectin 15 (SIGLEC15), T cell immunoreceptor with Ig and ITIM domains (TIGIT), CD274 Molecule (CD274), hepatitis A virus cellular receptor 2 (HAVCR2), programmed cell death 1 (PDCD1), cytotoxic T-lymphocyte associated protein 4 (CTLA4), lymphocyte activating 3 (LAG3) and programmed cell death 1 ligand 2 (PDCD1LG2) (Zeng et al., 2019).

Statistical Analysis

All the analytical methods (excluding functional enrichment analysis) were performed using Xiantao Academic (https://www.xiantao.love/products) embedded with R software and R packages, including org.Hs.eg.db, GEOquery, limma, ggplot2, clusterProfiler, survminer, survival and pROC (Davis and Meltzer, 2007; Yu et al., 2012; Liu et al., 2018; Hu et al., 2020). Chi-square test, paired *t* test and the Wilcoxon rank sum test were used to compare data. A value of *p* value < 0.05 was regarded as statistically significant.

RESULTS

Clinicopathological Characteristics

The flowchart of the present study is presented in **Figure 1**. RNA-seq expression data from 624 samples, including 174 normal tissues, 36 para carcinoma tissues and 414 tumor tissues plus clinical data were downloaded from UCSC XENA. The details are presented in **Table 2**.

TABLE 2 | Clinical characteristics of the GC patients based on TCGA.

Characteristic	Low expression of P4HA3	High expression of P4HA3	*p*
n	187	188	
T stage, n (%)			0.003
T1	17 (4.6%)	2 (0.5%)	
T2	41 (11.2%)	39 (10.6%)	
T3	85 (23.2%)	83 (22.6%)	
T4	43 (11.7%)	57 (15.5%)	
N stage, n (%)			0.885
N0	55 (15.4%)	56 (15.7%)	
N1	51 (14.3%)	46 (12.9%)	
N2	35 (9.8%)	40 (11.2%)	
N3	38 (10.6%)	36 (10.1%)	
M stage, n (%)			0.988
M0	166 (46.8%)	164 (46.2%)	
M1	12 (3.4%)	13 (3.7%)	
Pathologic stage, n (%)			0.154
Stage I	34 (9.7%)	19 (5.4%)	
Stage II	50 (14.2%)	61 (17.3%)	
Stage III	76 (21.6%)	74 (21%)	
Stage IV	19 (5.4%)	19 (5.4%)	
Primary therapy outcome, n (%)			0.099
PD	38 (12%)	27 (8.5%)	
SD	7 (2.2%)	10 (3.2%)	
PR	0 (0%)	4 (1.3%)	
CR	116 (36.6%)	115 (36.3%)	
Gender, n (%)			0.564

(Continued on following page)

TABLE 2 | (*Continued*) Clinical characteristics of the GC patients based on TCGA.

Characteristic	Low expression of P4HA3	High expression of P4HA3	*p*
Female	70 (18.7%)	64 (17.1%)	
Male	117 (31.2%)	124 (33.1%)	
Age, n (%)			0.437
≤65	86 (23.2%)	78 (21%)	
>65	99 (26.7%)	108 (29.1%)	
Histological type, n (%)			0.005
Diffuse Type	27 (7.2%)	36 (9.6%)	
Mucinous Type	7 (1.9%)	12 (3.2%)	
Not Otherwise Specified	97 (25.9%)	110 (29.4%)	
Papillary Type	3 (0.8%)	2 (0.5%)	
Signet Ring Type	4 (1.1%)	7 (1.9%)	
Tubular Type	49 (13.1%)	20 (5.3%)	
Histologic grade, n (%)			0.009
G1	5 (1.4%)	5 (1.4%)	
G2	82 (22.4%)	55 (15%)	
G3	95 (26%)	124 (33.9%)	
H pylori infection, n (%)			0.869
No	88 (54%)	57 (35%)	
Yes	10 (6.1%)	8 (4.9%)	
Barretts esophagus, n (%)			0.461
No	116 (55.8%)	77 (37%)	
Yes	11 (5.3%)	4 (1.9%)	

Abbreviations: CR, complete response; PD, progressive disease; SD, stable disease; PR, partial response.

Expression of P4HA3 Across Many Cancer Cell-Types

Differential expression of P4HA3 mRNA was measured and found to be over-expressed in 16 cancers, including breast invasive carcinoma (BRCA), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), acute myeloid leukemia (LAML), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), rectum adenocarcinoma (READ), thymoma (THYM) and stomach adenocarcinoma (STAD; **Figure 2A**). All-sample and paired sample analysis showed that expression levels of P4HA3 were higher in GC tissue than in non-cancerous tissue (**Figures 2B,C**). Further analysis using GSE54129 (**Figure 2D**) and GSE103236 (**Figure 2E**) gave similar results to those from the TCGA data.

P4HA3 Expression and GC Clinicopathological Features

The correlation analysis showed significant differences for some clinicopathological features, including *Helicobacter pylori* infection status (**Figure 3A**), pathological stage (**Figure 3B**), T classification (**Figure 3C**) and histological grade (**Figure 3D**). There were no differences in P4HA3 expression based on gender (**Figure 3E**) or age (**Figure 3F**).

Correlation Analysis of Prognosis

The area under the ROC curve was 0.933, based on TCGA data (**Figure 4A**), and 0.874 for non-cancerous samples of GTEx combined para carcinoma tissues and GC samples (**Figure 4B**). These results indicate that levels of P4HA3 expression are consistently different in tumor and non-tumor tissues. Kaplan-Meier survival analysis indicated that high levels of P4HA3 are associated with poor prognosis (**Figure 4C**).

Univariate analysis demonstrated that high P4HA3 expression corresponded to reduced OS and, thus, poor prognosis for GC patients (**Table 3**). TNM stage, pathological stage and age were also associated with reduced OS (**Table 3**). The results of multivariate analysis showed that P4HA3 was an independent prognostic marker. TNM stage and age also had independent prognostic value for OS in GC (**Table 3**).

PPI Networks and Enrichment Analysis

The present study reports the construction of a network of P4HA3 and its related genes using the STRING database. P4HA3-related genes with scores above 0.9 included Collagen Type I Alpha (COL1A)1, COL1A2, COL3A1, COL6A3, COL12A1, COL20A1, prolyl 3-hydroxylase (P3H)1, P3H2 and P3H3 (**Figure 5**; **Table 4**). Metascape pathway and process enrichment analysis revealed that all genes related to P4HA3 were enriched in R-HSA-1650814, suggesting roles in collagen biosynthesis and modifying enzymes. The endoplasmic reticulum lumen and procollagen-proline 3-dioxygenase activity were also associated with P4HA3 (**Table 5**).

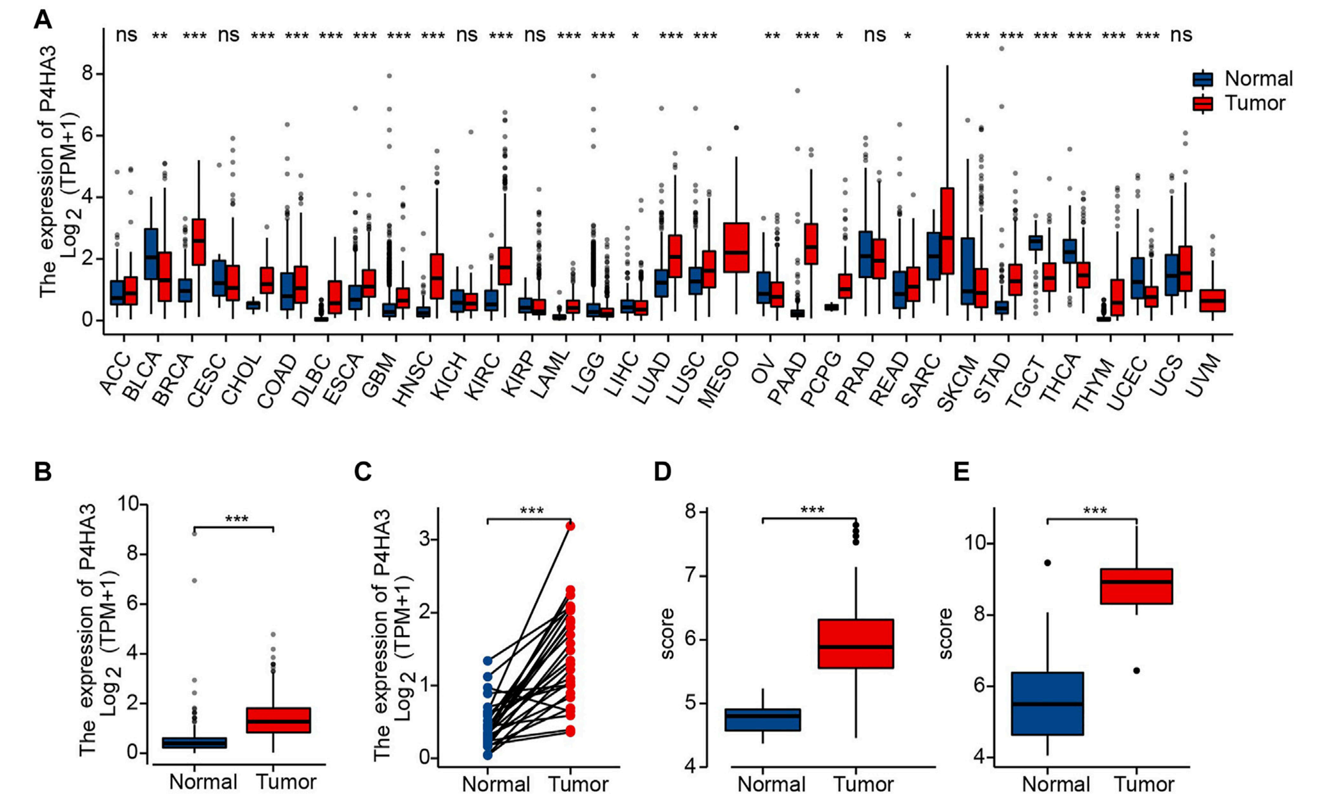


FIGURE 2 | Upregulation of P4HA3 in GC. **(A)** P4HA3 expression levels in various cancer-types from TCGA data; **(B)** P4HA3 transcript levels in GC and non-cancerous gastric tissues from TCGA data; **(C)** P4HA3 expression in paired samples; **(D)** P4HA3 expression in GSE54129; **(E)** P4HA3 expression in GSE103236. (ns, $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.) Abbreviations: P4HA3, prolyl 4-hydroxylase subunit alpha 3; GC, gastric cancer; BRCA, breast invasive carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; DLBC, lymphoid neoplasm diffuse large B-cell lymphoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme, HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; LAML, acute myeloid leukemia; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; PAAD, pancreatic adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; READ, rectum adenocarcinoma; THYM, thymoma; STAD, stomach adenocarcinoma; THCA, thyroid carcinoma.

Correlation Analysis of Immune Infiltration

The association between P4HA3 and the degree of immune cell infiltration in GC was explored using ssGSEA analysis (**Figure 6A**). Macrophages, NK cells, iDC, mast cells, eosinophils, Tem, Th1 cells and DC all correlated with P4HA3 (**Figures 6B–I**).

Patients were divided into two groups according to P4HA3 expression and those with high expression had higher levels of the immune-checkpoint–relevant transcripts, HAVCR2 and PDCD1LG2, than those with low expression (**Figure 7A**). P4HA3 expression was positively correlated with HAVCR2 (**Figure 7B**) and PDCD1LG2 (**Figure 7C**).

DISCUSSION

Prolyl 4-hydroxylase (P4H) activity is essential for maintenance of the collagen triple helix and P4HAs (P4HA1, P4HA2, P4HA3) plus P4HB are highly expressed in numerous tumors where they may contribute to cancer progression. A number of inhibitors of P4HAS and P4HB have been shown to exert anti-tumor effects, suggesting that P4H is an achievable target for cancer therapy (Shi et al., 2021). Expression profiles and functional roles of P4HA3 in GC have rarely been studied. The current study focused on the diagnostic, prognostic and potential immune therapeutic target value of P4HA3 in GC.

mRNA expression levels across many different cancer types were analyzed using data from the TCGA database. P4HA3 mRNA was up-regulated in GC, along with BRCA, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KIRC, LAML, LUAD, LUSC, PAAD, PCPG, READ and THYM, in agreement with the previous study of Hu et al., 2020. P4HA3 up-regulation was confirmed using GSE54129 and GSE103236 from the GEO database.

A key component of the current study was to address the diagnostic and prognostic value of P4HA3. GC patients tended to have higher P4HA3 mRNA expression when they were infected by HP, resulting in diagnosis with higher T stages and lower histological grades, or when they were diagnosed at late pathological stages. ROC analysis indicated differences in P4HA3 expression between tumor tissues and non-cancerous tissues. Kaplan–Meier survival analysis indicated that those GC patients with higher levels of P4HA3 tended to have shorter OS. Multivariate Cox analysis demonstrated that high levels of P4HA3 mRNA constituted an independent risk factor for OS

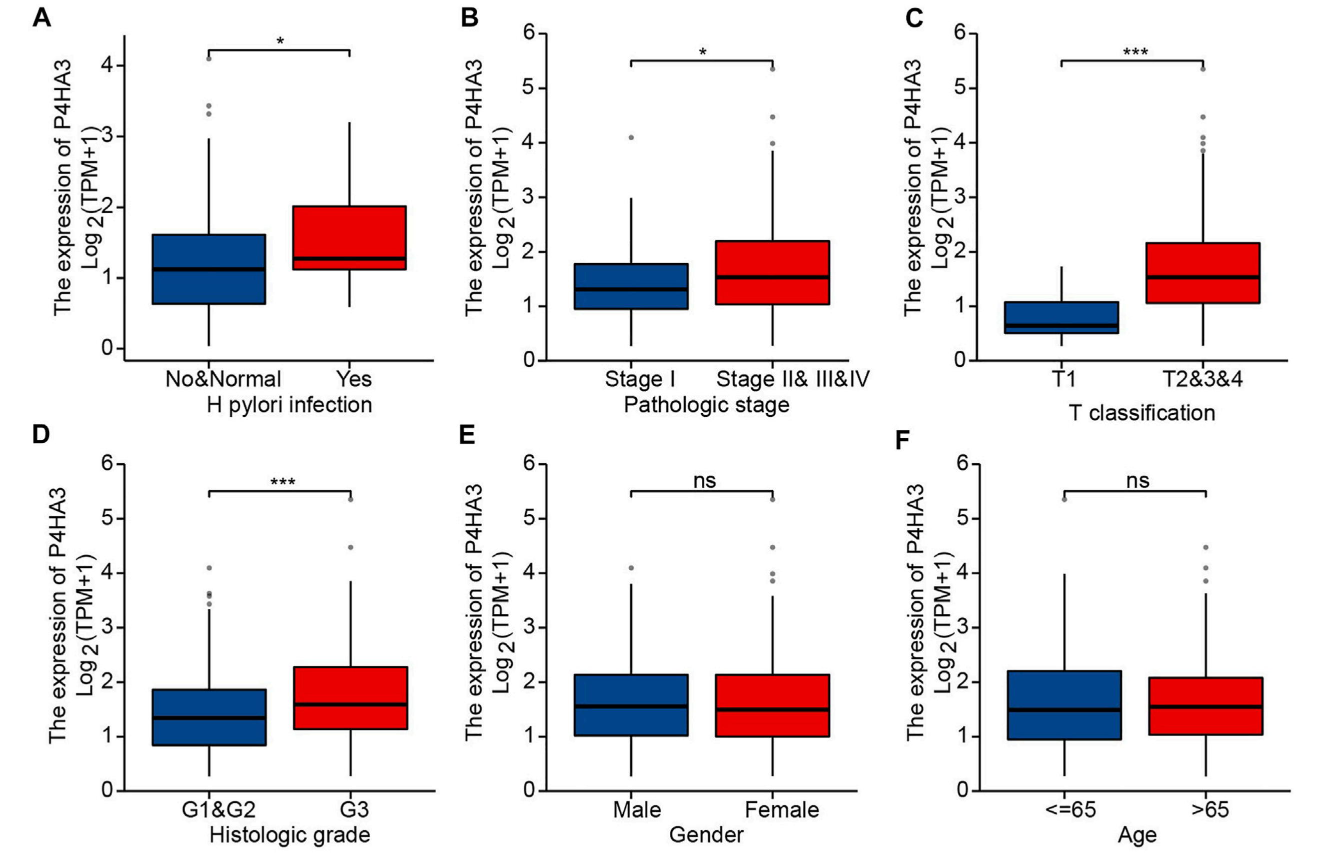


FIGURE 3 | P4HA3 expression is associated with clinicopathological characteristics. **(A)** Post-mortal P4HA3 expression levels were higher; **(B)** Late stage P4HA3 expression levels were higher; **(C)** P4HA3 expression levels were higher in patients with higher T classifications; **(D)** P4HA3 expression levels were higher in patients with lower histological grades; **(E)** No gender differences were found for P4HA3 expression levels; **(F)** No age-related changes were found in P4HA3 expression levels. ns, $p \geq 0.05$; *, $p < 0.05$; ***, $p < 0.001$.

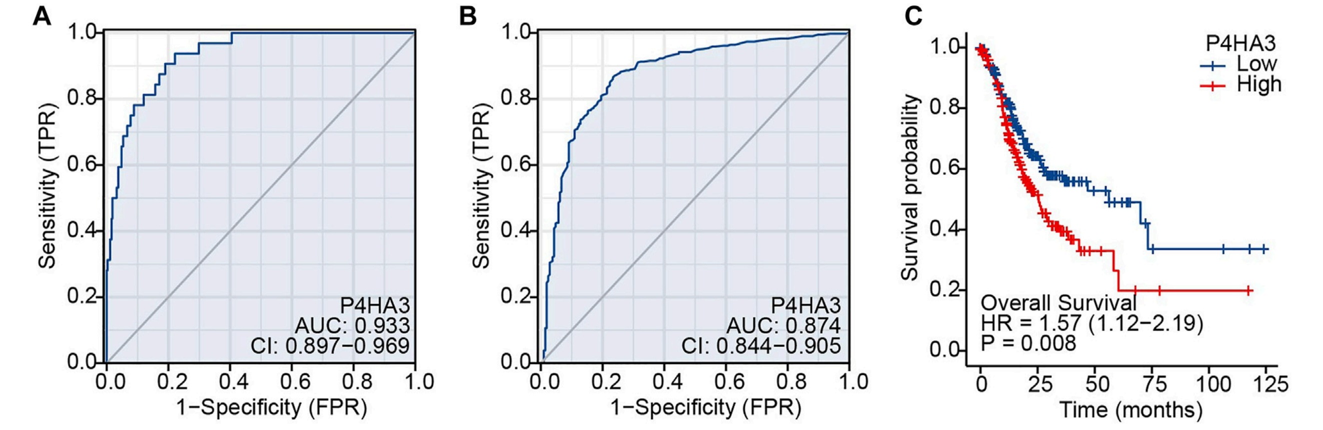


FIGURE 4 | ROC and Kaplan-Meier survival curves. **(A)** ROC curve for P4HA3 based on data from TCGA; **(B)** ROC curve for P4HA3 using data from non-cancerous samples from GTEx combined para carcinoma tissues and GC samples; **(C)** Higher levels of P4HA3 expression tended to be associated with worse outcomes (OS) for GC patients. Abbreviations: ROC, receiver operating characteristic; GTEx, Genotype-Tissue Expression Project.

and were associated with poor GC prognosis. Thus, we believe that P4HA3 could serve as a novel diagnostic and independent prognostic biomarker for GC patients.

P4HA3 has been shown to have an association with many different cancers. It promoted cell proliferation, invasion and migration in head and neck squamous cell carcinoma and

TABLE 3 | Univariate and multivariate Cox regression analysis of the P4HA3 expression and overall survival in gastric cancer patients.

Variable	Total (*N*)	Univariate analysis		Multivariable	
		HR (95% CI)	*p* value	HR (95% CI)	*p* value
T stage (T1 vs. T2&3&4)	362	8.829 (1.234-63.151)	**0.030**	3.735 (0.502-27.792)	0.198
N stage (N0 vs. N1&2&3)	352	1.925 (1.264-2.931)	**0.002**	1.356 (0.749-2.454)	0.314
M stage (M0 vs. M1)	352	2.254 (1.295-3.924)	**0.004**	1.959 (1.015-3.781)	**0.045**
Pathological stage (Stage I& II vs. III& IV)	347	1.947 (1.358-2.793)	**<0.001**	1.371 (0.819-2.294)	0.230
Grade (G1&G2 vs. G3)	361	1.353 (0.957-1.914)	0.087	1.290 (0.878-1.894)	0.194
Age (≤65 vs. >65)	367	1.620 (1.154-2.276)	**0.005**	1.866 (1.278-2.725)	**0.001**
Gender (Female vs. Male)	370	1.267 (0.891-1.804)	0.188	-	-
H pylori infection (NO vs. Yes)	162	0.650 (0.279-1.513)	0.317	-	-
Reflux history (No vs. Yes)	213	0.582 (0.291-1.162)	0.125	-	-
P4HA3 (Low vs. High)	370	1.634 (1.169-2.284)	**0.004**	1.641 (1.135-2.374)	**0.008**

Notes: Bold type indicates statistical significance.
Abbreviations: P4HA3, Prolyl 4-hydroxylase subunit alpha 3; HR, hazard ratio; CI: confidence interval.

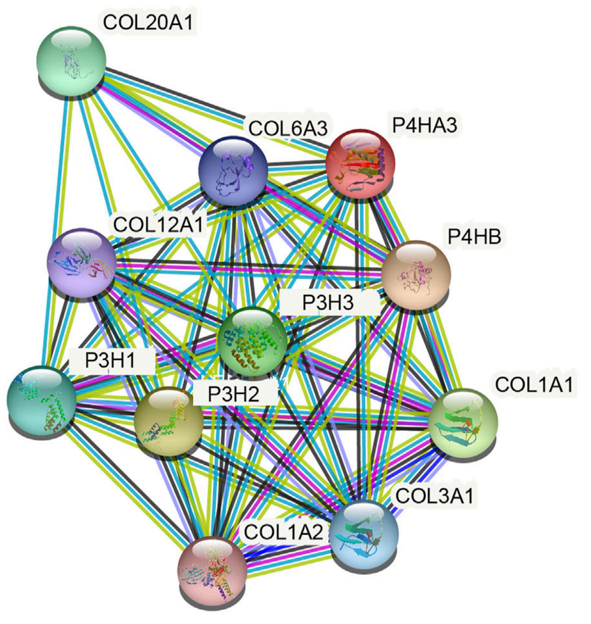


FIGURE 5 | Protein-protein interaction (PPI) network of P4HA3-related proteins (TOP 10, medium confidence).

TABLE 4 | The detailed information of P4HA3-related genes.

Gene symbol	Annotation	Score
COL1A1	collagen type I alpha 1 chain	0.951
COL1A2	collagen type I alpha 2 chain	0.926
COL3A1	collagen type III alpha 1 chain	0.933
COL6A3	collagen type VI alpha 3 chain	0.927
COL12A1	collagen type XII alpha 1 chain	0.927
COL20A1	collagen type XX alpha 1 chain	0.941
P3H1	prolyl 3-hydroxylase 1	0.934
P3H2	prolyl 3-hydroxylase 2	0.959
P3H3	prolyl 3-hydroxylase 3	0.944

melanoma cells (Atkinson et al., 2019; Wang et al., 2020) and reduced the anti-tumor activity of COL6A6 on growth and metastasis of AtT-20 and HP75 melanoma cells due to an action on PI3K-Akt signaling (Long et al., 2019). P4HA3 is known to be upregulated in clear cell renal carcinoma and patients with higher expression had worse outcomes, indicating a prognostic role for P4HA3 (Liu et al., 2020). Findings of the present and previous studies indicate that P4HA3 maintains the stability of newly synthesized collagens and remodels the extracellular matrix in GC (Shoulders and Raines, 2009; Gilkes et al., 2013, 2014).

Macrophages and NK cells influence the tumor microenvironment and tumor immunity (Lee et al., 2014; Sammarco et al., 2019; Gambardella et al., 2020). Infiltration of M2 macrophages promotes tumor cell escape and thus, numbers may reflect prognosis (Liu X. et al., 2021). Infiltration of NK cells has an impact on immunotherapy and targeting NK cells may improve the anti-tumor immune response (Yang et al., 2019; Bi et al., 2021). P4HA3 expression correlated with immune infiltration by macrophages and NK cells.

Immune checkpoint molecules regulate self-tolerance to prevent autoimmune reactions and minimize tissue damage by controlling the length and intensity of the immune response. Expression of checkpoint molecules acts to limit the immune and anti-tumor immune response, enabling escape of tumor cells (Galluzzi et al., 2020; Liu Y. et al., 2021). Patients with higher P4HA3 mRNA expression tended to have higher expression of the immune checkpoint related genes, PDCD1LG2 and HAVCR2. The current findings indicate the potential for targeting of P4HA3 during GC immunotherapy.

In conclusion, the purpose of the current study was to determine the diagnostic and prognostic value of P4HA3 and its correlation with immune cell infiltration in GC. P4HA3 emerges as a feasible diagnostic and prognostic biomarker and immunotherapy target. However, the current results are all derived from bioinformatics analysis and limited by the absence of experimental confirmation. Further clinical experiments are underway to verify the function of P4HA3 in GC.

TABLE 5 | Clusters with their representative enriched terms (one per cluster).

Term	Category	Description	Count (%)	*p*	q
R-HSA-1650814	Reactome Gene Sets	Collagen biosynthesis and modifying enzymes	9 (100)	<0.001	<0.001
M3005	Canonical Pathways	NABA COLLAGENS	6 (66.67)	<0.001	<0.001
GO:0005788	GO Cellular Components	endoplasmic reticulum lumen	8 (88.89)	<0.001	<0.001
GO:0019797	GO Molecular Functions	procollagen-proline 3-dioxygenase activity	3 (33.33)	<0.001	<0.001

Abbreviation: GO, gene ontology.

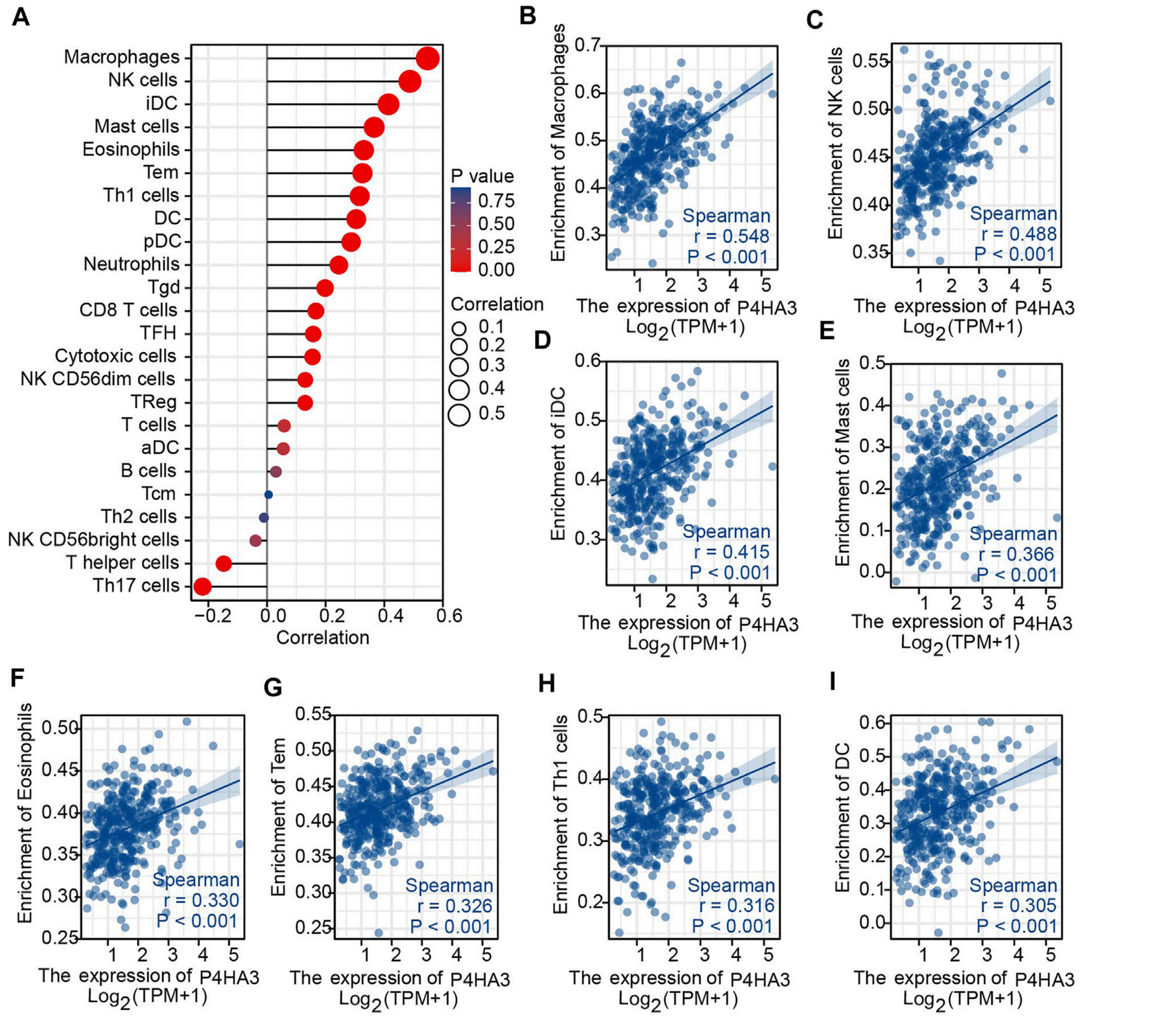


FIGURE 6 | Immune Cell Infiltration Analysis. **(A)** The Lollipop Chart shows the correlation between P4HA3 expression level and 24 different immune cell-types; **(B–I)** the enrichment scores of P4HA3 expression for 8 immune cell-types. Abbreviations: NK, natural killer; iDC, immature dendritic cells; Tem, effective memory T cells; Th1, T-helper 1, DC, dendritic cells.

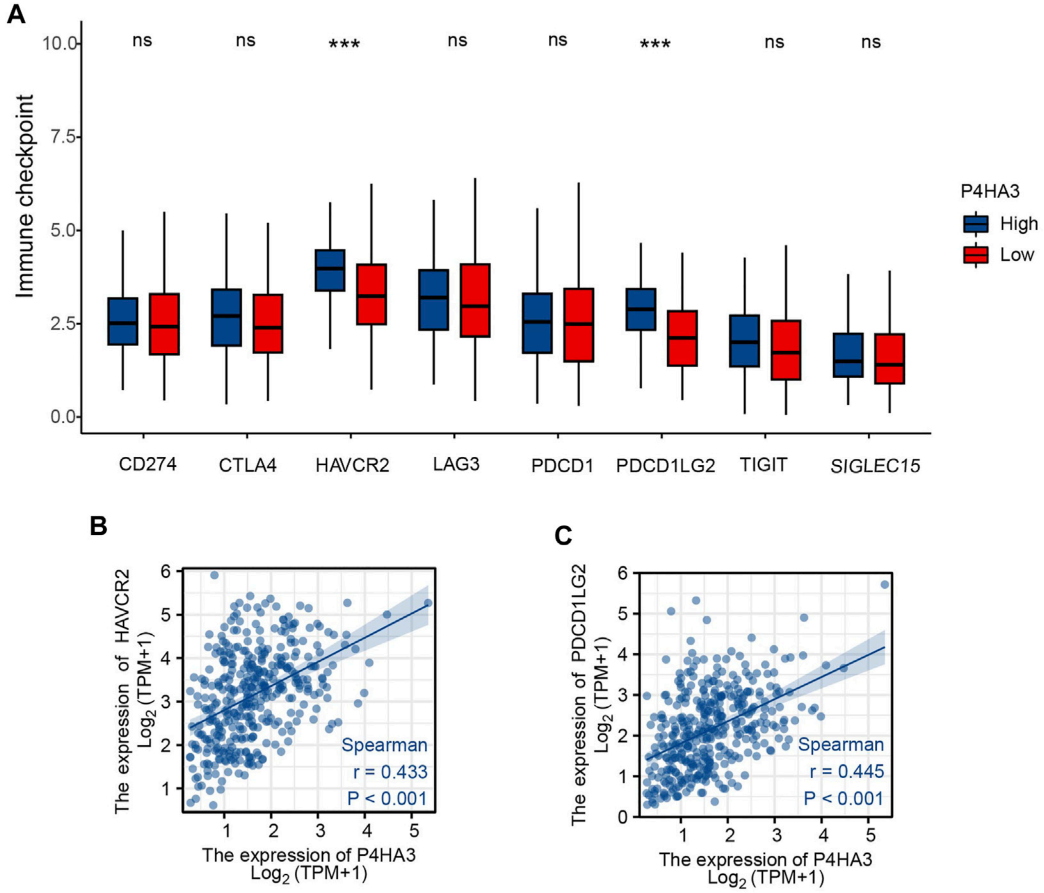


FIGURE 7 | (A) Differential expression of immune-checkpoint–relevant genes in low and high P4HA3-expressing groups. **(B)** P4HA3 was positively correlated with HAVCR2. **(C)** P4HA3 was positively correlated with PDCD1LG2. ns, $p > 0.05$; ***, $p < 0.001$. Abbreviations: HAVCR2, hepatitis A virus cellular receptor 2; PDCD1LG2, programmed cell death 1 ligand 2.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

XN, SZ, and HQ contributed to conception and design of the study. XN, LR, and SW organized the database and performed the statistical analysis. XN, LR, and SW wrote the first draft of the manuscript. DG, MM, and AH wrote sections of the manuscript. SZ and HQ contributed to manuscript revision. All authors read and approved the submitted version.

FUNDING

The present work was funded by the Construction Project of Provincial Clinical Key Subject from the Health Commission of Qinghai Province.

ACKNOWLEDGMENTS

We would like to thank TCGA and GEO databases for providing platforms and contributors for uploading their meaningful data sets. The authors thank Xiantao Academic for their contribution in data analyzing and data visualizing. The present study proposal was reviewed by an institutional board, Ethics Committee of Qinghai Province Hospital of Traditional Chinese Medicine, and that the need for ethics approval had been waived.

REFERENCES

Atkinson, A., Renziehausen, A., Wang, H., Lo Nigro, C., Lattanzio, L., Merlano, M., et al. (2019). Collagen Prolyl Hydroxylases Are Bifunctional Growth Regulators in Melanoma. *J. Investigative Dermatology* 139, 1118–1126. doi:10.1016/j.jid.2018.10.038

Bi, J., Cheng, C., Zheng, C., Huang, C., Zheng, X., Wan, X., et al. (2021). TIPE2 Is a Checkpoint of Natural Killer Cell Maturation and Antitumor Immunity. *Sci. Adv.* 7, eabi6515. eabi6515. doi:10.1126/sciadv.abi6515

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal Dynamics of Intratumoral Immune Cells Reveal the Immune Landscape in Human Cancer. *Immunity* 39, 782–795. doi:10.1016/j.immuni.2013.10.003

Davis, S., and Meltzer, P. S. (2007). GEOquery: a Bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* 23, 1846–1847. doi:10.1093/bioinformatics/btm254

Galluzzi, L., Humeau, J., Buqué, A., Zitvogel, L., and Kroemer, G. (2020). Immunostimulation with Chemotherapy in the Era of Immune Checkpoint Inhibitors. *Nat. Rev. Clin. Oncol.* 17, 725–741. doi:10.1038/s41571-020-0413-z

Gambardella, V., Castillo, J., Tarazona, N., Gimeno-Valiente, F., Martínez-Ciarpaglini, C., Cabeza-Segura, M., et al. (2020). The Role of Tumor-Associated Macrophages in Gastric Cancer Development and Their Potential as a Therapeutic Target. *Cancer Treat. Rev.* 86, 102015. doi:10.1016/j.ctrv.2020.102015

Gilkes, D. M., Chaturvedi, P., Bajpai, S., Wong, C. C., Wei, H., Pitcairn, S., et al. (2013). Collagen Prolyl Hydroxylases Are Essential for Breast Cancer Metastasis. *Cancer Res.* 73, 3285–3296. doi:10.1158/0008-5472.CAN-12-3963

Gilkes, D. M., Semenza, G. L., and Wirtz, D. (2014). Hypoxia and the Extracellular Matrix: Drivers of Tumour Metastasis. *Nat. Rev. Cancer* 14, 430–439. doi:10.1038/nrc3726

Gu, X., Wang, B., Zhu, H., Zhou, Y., Horning, A. M., Huang, T. H.-M., et al. (2020). Age-associated Genes in Human Mammary Gland Drive Human Breast Cancer Progression. *Breast Cancer Res.* 22, 64. doi:10.1186/s13058-020-01299-2

Güthle, M., Ettrich, T., and Seufferlein, T. (2020). Immunotherapy in Gastrointestinal Cancers. *Visc. Med.* 36, 231–237. doi:10.1159/000507798

Hochberg, Y., and Benjamini, Y. (1990). More Powerful Procedures for Multiple Significance Testing. *Stat. Med.* 9, 811–818. doi:10.1002/sim.4780090710

Hu, W., Wang, G., Chen, Y., Yarmus, L. B., Liu, B., and Wan, Y. (2020). Coupled Immune Stratification and Identification of Therapeutic Candidates in Patients with Lung Adenocarcinoma. *Aging* 12, 16514–16538. doi:10.18632/aging.103775

Lee, K., Hwang, H., and Nam, K. T. (2014). Immune Response and the Tumor Microenvironment: How They Communicate to Regulate Gastric Cancer. *Gut Liver* 8, 131–139. doi:10.5009/gnl.2014.8.2.131

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Kovatich, A. D. A. J., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* 173, 400–e11. e11. doi:10.1016/j.cell.2018.02.052

Liu, M., Pan, Q., Xiao, R., Yu, Y., Lu, W., and Wang, L. (2020). A Cluster of Metabolism-Related Genes Predict Prognosis and Progression of Clear Cell Renal Cell Carcinoma. *Sci. Rep.* 10, 12949. doi:10.1038/s41598-020-67760-6

Liu, X., Hogg, G. D., and DeNardo, D. G. (2021a). Rethinking Immune Checkpoint Blockade: 'Beyond the T Cell'. *J. Immunother. Cancer* 9, e001460. doi:10.1136/jitc-2020-001460

Liu, Y., Xiang, H., Zhu, J., and Fang, Z. (2021b). Progress of Tumor-Associated Macrophages in Tumors. *Cancer Res. Clin.* 33, 149–153. doi:10.3760/cma.j.cn115355-20191230-00598

Long, R., Liu, Z., Li, J., and Yu, H. (2019). COL6A6 Interacted with P4HA3 to Suppress the Growth and Metastasis of Pituitary Adenoma via Blocking PI3K-Akt Pathway. *Aging* 11, 8845–8859. doi:10.18632/aging.102300

Luo, Y., Xu, W., Chen, H., Warburton, D., Dong, R., Qian, B., et al. (2015). A Novel Profibrotic Mechanism Mediated by TGFβ-Stimulated Collagen Prolyl Hydroxylase Expression in Fibrotic Lung Mesenchymal Cells. *J. Pathol.* 236, 384–394. doi:10.1002/path.4530

Machlowska, J., Baj, J., Sitarz, M., Maciejewski, R., and Sitarz, R. (2020). Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies. *Ijms* 21, 4012. doi:10.3390/ijms21114012

Niu, X., Ren, L., Hu, A., Zhang, S., and Qi, H. (2022). Identification of Potential Diagnostic and Prognostic Biomarkers for Gastric Cancer Based on Bioinformatic Analysis. *Front. Genet.* 13, 862105. doi:10.3389/fgene.2022.862105

Sammarco, G., Varricchi, G., Ferraro, V., Ammendola, M., De Fazio, M., Altomare, D. F., et al. (2019). Mast Cells, Angiogenesis and Lymphangiogenesis in Human Gastric Cancer. *Ijms* 20, 2106. doi:10.3390/ijms20092106

Shi, R., Gao, S., Zhang, J., Xu, J., Graham, L. M., Yang, X., et al. (2021). Collagen Prolyl 4-hydroxylases Modify Tumor Progression. *Acta Biochim. Biophys. Sin. (Shanghai).* 53, 805–814. doi:10.1093/abbs/gmab065

Shoulders, M. D., and Raines, R. T. (2009). Collagen Structure and Stability. *Annu. Rev. Biochem.* 78, 929–958. doi:10.1146/annurev.biochem.77.032207.120833

Song, H., Liu, L., Song, Z., Ren, Y., Li, C., and Huo, J. (2018). P4HA3is Epigenetically Activated by Slug in Gastric Cancer and its Deregulation Is Associated with Enhanced Metastasis and Poor Survival. *Technol. Cancer Res. Treat.* 17, 153303381879648. doi:10.1177/1533033818796485

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA A Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., et al. (2021). The STRING Database in 2021: Customizable Protein-Protein Networks, and Functional Characterization of User-Uploaded Gene/measurement Sets. *Nucleic Acids Res.* 49, D605–D612. doi:10.1093/nar/gkaa1074

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., et al. (2017). Toil Enables Reproducible, Open Source, Big Biomedical Data Analyses. *Nat. Biotechnol.* 35, 314–316. doi:10.1038/nbt.3772

Wang, T., Wang, Y.-X., Dong, Y.-Q., Yu, Y.-L., and Ma, K. (2020). Prolyl 4-hydroxylase Subunit Alpha 3 Presents a Cancer Promotive Function in Head and Neck Squamous Cell Carcinoma via Regulating Epithelial-Mesenchymal Transition. *Archives Oral Biol.* 113, 104711. doi:10.1016/j.archoralbio.2020.104711

Yang, C-M., Zhang, T-t., Cheng, Q., Zou, W., Wen-xing, C., Wang, A-y., et al. (2019). Progress in Research on Anti-tumor Mechanism of NK Cell and its Related Immunotherapy. *Chin. Pharmacol. Bull.* 35, 1492–1495. doi:10.3969/j.issn.1001-1978.2019.11.003

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS A J. Integr. Biol.* 16, 284–287. doi:10.1089/omi.2011.0118

Zeng, D., Li, M., Zhou, R., Zhang, J., Sun, H., Shi, M., et al. (2019). Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures. *Cancer Immunol. Res.* 7, 737–750. doi:10.1158/2326-6066.CIR-18-0436

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape Provides a Biologist-Oriented Resource for the Analysis of Systems-Level Datasets. *Nat. Commun.* 10, 1523. doi:10.1038/s41467-019-09234-6

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

TYPE Original Research
PUBLISHED 09 September 2022
DOI 10.3389/fgene.2022.984575

Check for updates

OPEN ACCESS

EDITED BY
Saurav Mallik,
Harvard University, United States

REVIEWED BY
Loveleen Gaur,
Amity University, India
Soumita Seth,
Aliah University, India

*CORRESPONDENCE
Hua Yan,
71001643@sdutcm.edu.cn

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 02 July 2022
ACCEPTED 08 August 2022
PUBLISHED 09 September 2022

CITATION
Wei Q, Miao T, Zhang P, Jiang B and Yan H (2022), Comprehensive analysis to identify GNG7 as a prognostic biomarker in lung adenocarcinoma correlating with immune infiltrates. *Front. Genet.* 13:984575. doi: 10.3389/fgene.2022.984575

Comprehensive analysis to identify GNG7 as a prognostic biomarker in lung adenocarcinoma correlating with immune infiltrates

Qin Wei[1], Tianshu Miao[1], Pengju Zhang[1], Baodong Jiang[2] and Hua Yan[3]*

[1]Department of Biochemistry and Molecular Biology, Shandong University School of Basic Medical Sciences, Jinan, China, [2]Department of Radiology, Qilu Hospital of Shandong University, Jinan, China, [3]Department of Gastroenterology, Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan, China

Background: G Protein Subunit Gamma 7 (GNG7), an important regulator of cell proliferation and cell apoptosis, has been reported to be downregulated in a variety of tumors including lung adenocarcinoma (LUAD). However, the correlation between low expression of GNG7 and prognosis of LUAD as well as the immune infiltrates of LUAD remains unclear.

Methods: The samples were obtained from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). R software was performed for statistical analysis. GNG7 expression and its prognostic value in LUAD were assessed through statistically analyzing the data from different databases. A nomogram was constructed to predict the impact of GNG7 on prognosis. Gene set enrichment analysis (GSEA) and single-sample gene set enrichment analyses GSEA (ssGSEA) were employed to determine the potential signal pathways and evaluated the immune cell infiltration regulated by GNG7. The prognostic significance of GNG7 expression associated with immune cell infiltration was investigated using the Tumor Immune Estimation Resource 2.0 (TIMER2.0) and the Kaplan-Meier plotter database. The UALCAN, cBio Cancer Genomics Portal (cBioPortal) and MethSurv database were used to analyze the correlation between the methylation of GNG7 and its mRNA expression as well as prognostic significance.

Results: GNG7 was demonstrated to be down-regulated in LUAD and its low expression was associated with poor prognosis. A clinical reliable prognostic-predictive model was constructed. Pathway enrichment showed that GNG7 was highly related to the B cell receptor signaling pathway. Further analysis showed that GNG7 was positively associated with B cell infiltration and low levels of B cell infiltration tended to associate with worse prognosis in patients with low GNG7 expression. Moreover, methylation analysis suggested hypermethylation may contribute to the low expression of GNG7 in LUAD.

Conclusion: Decreased expression of GNG7 at least partly caused by hypermethylation of the GNG7 promoter is closely associated with poor prognosis and tumor immune cell infiltration (especially B cells) in LUAD. These results suggest that GNG7 may be a promising prognostic biomarker and a potential immunotherapeutic target for LUAD, which provides new insights into immunotherapy for LUAD.

KEYWORDS

immune microenvironment 1, GNG7 2, prognosis 3, lung adenocarcinoma 4, bioinformatics analysis 5

Introduction

Lung cancer is the most common reason for global cancer-related mortality, of which lung adenocarcinoma (LUAD) is the most common histological subtype (Travis, 2011; Sung et al., 2021). In recent years, although molecular targeted therapies and immunotherapy have significantly improved the prognosis of a small proportion of LUAD patients, the above treatments are ineffective in many patients due to the high heterogeneity and complexity of LUAD (Molina et al., 2008; Saito et al., 2018; Wu and Shih, 2018). The prognosis for many patients, especially those with advanced LUAD, remains poor, with a 5-year survival rate of less than 18% (Singh et al., 2020). Therefore, an in-depth pathogenetic exploration and search for other effective diagnostic and therapeutic approaches as well prognostic markers are essential to improve the prognosis of patients with LUAD.

Accumulating evidence suggests that the immune cells within the tumor microenvironment (TME) play essential roles in tumorigenesis (Hinshaw and Shevde, 2019). Such tumor associated immune cells may exert pro-tumor or anti-tumor function in the initiation and development of tumors (Taube et al., 2018). Studies have shown that immune cell infiltration is an important factor influencing the efficacy of immunotherapy (Martinez and Moon, 2019; Petitprez et al., 2020; Bagchi et al., 2021). In addition, TME is also closely related to the prognosis of patients (Qi et al., 2020; Wu et al., 2020). Therefore, it is crucial to investigate the regulators of immune cell infiltration in the TME to improve the effectiveness of immunotherapy and improve patient prognosis.

G Protein Subunit Gamma 7 (GNG7), a subunit of heterotrimeric G protein, is strongly enriched in the striatum and plays a vital role in the A2A adenosine or D1 dopamine receptor-induced neuro-protective response (Schwindinger et al., 2012). Multiple studies have shown that GNG7 expression is decreased in many cancers, including pancreatic cancer, gastrointestinal tract cancer and renal carcinoma (Shibata et al., 1998; Shibata et al., 1999; Ohta et al., 2008). GNG7 overexpression was shown to inhibit cell growth and tumorigenicity of esophageal carcinoma cells (Hartmann et al., 2012). Also, GNG7 was confirmed as an essential autophagy-inducing agent and participated in inhibiting tumor progression through mTOR pathway (Xu et al., 2019). Recently, GNG7 was reported to be lowly expressed in LUAD and promoted the progression of LUAD through Hedgehog signaling (Liu et al., 2016). These findings indicated that GNG7 may be a potential tumor suppressor implicated in the carcinogenesis and tumor progression. However, the detailed roles and mechanisms of GNG7 especially the effects of GNG7 on immune infiltration in LUAD are largely unknown.

In this study, we aimed to evaluate the clinical significance of GNG7 in LUAD and the possible mechanisms underlying its function through comprehensive bioinformatics analysis. Our results showed that low expression of GNG7 positively correlated with the progression of LUAD, and GNG7 may be an important potential prognostic biomarker for LUAD. We also constructed a reliable clinical prediction model. In addition, we revealed for the first time the underlying mechanisms of GNG7 dysregulation and the correlation between GNG7 expression and immune infiltration in LUAD, which implies that GNG7 could be a potential target for clinical antitumor immunotherapy.

Materials and methods

Data sources and pretreatment

The RNA-seq data of 513 LUAD samples and 59 normal samples were downloaded from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/). The downloaded data format was level 3 HTSeq-fragments per kilobase per million (FPKM) and then was converted to transcripts per million (TPM) format for subsequent analysis. TCGA supplemented prognostic data were obtained from a Cell article (Liu et al., 2018). In addition, three sets of microarray data of LUAD tissues (accession numbers: GSE32665, GSE32863, and GSE43458) were downloaded from the GEO database. The data used in this study were obtained from both the TCGA database and the GEO database, which ensured that all written informed consent was obtained prior to data collection.

Key gene screening

The R package "DESeq2" was used to identify the differentially expressed genes (DEGs) between LUAD tissues and normal tissues (Love et al., 2014). Adjusted *p*-value <0.05 and | Log2fold change| >1 were set as cut-off criteria. The Survminer R package and the survivor R package were used to screen for genes with better prognostic value. The cut-off threshold was hazard ratio (HR) < 1, and the Cox *p*-value <0.005. The prognostic indicators included overall survival (OS), disease-specific survival (DSS) and progression-free survival (PFS). The Venn diagram was used to represent the intersection set of DEGs obtained from the four sets.

G protein subunit gamma 7 differential expression analysis in lung adenocarcinoma tissues

Pre-processed TCGA-LUAD data were used for differential expression analysis of GNG7 in LUAD tumor tissues and normal tissues. Three GEO expression profile datasets, GSE32665, GSE32863, and GSE43458, were used to compare the expression of GNG7 between LUAD tissues and normal tissues. Differential protein levels of GNG7 between LUAD tissues and normal tissues were analyzed by UALCAN (http://ualcan.path.uab.edu/) (Chandrashekar et al., 2022). A receiver operating characteristic (ROC) curve was used to evaluate the diagnostic significance of GNG7 using the plotROC R package (Version 1.17.0.1) (Robin et al., 2011).

Clinical statistical analysis on prognosis, model construction and evaluation

The correlation between GNG7 expression and survival in LUAD patients was analyzed by the PrognoScan database (http://www.prognoscan.org/) (Mizuno et al., 2009). Univariate Cox regression analysis, multivariate Cox regression analysis, logistic analysis and Kaplan-Meier (K-M) analysis were employed for prognosis analysis. The independent prognostic factors obtained from multivariate Cox regression analysis were employed to establish nomograms to predict survival probability for 1-, 2-, and 3-year. The calibration curves and nomograms were analyzed and plotted *via* the rms (version 6.2-0) and survival (version 3.2-10) package of R software. The calibration curves were graphically assessed by mapping the nomogram-predicted probabilities against the observed rates, and the 45-degree line represented the best predictive values. A concordance index (C-index) was used to determine the discrimination of the nomogram. According to the median risk score, patients were divided into a high-risk score group and a low-risk score group. The survival difference between the two groups was assessed by K-M survival curves. The model was compared with the two-by-two model consisting of independent prognostic factors screened from multivariate Cox regression, and ROC curves made by the timeROC R package (version 0.4) were used to assess the accuracy of the model predictions. The risk curve was used to demonstrate the application of the model in predicting clinical outcomes.

Gene set enrichment analysis

DESeq2 package (Version 1.26.0) was employed to identify the DEGs between GNG7-high and GNG7-low expression patients from TCGA datasets. The cut-off threshold was |log fold change (FC)|>1 and adjusted *p*-value <0.05. All the DEGs were presented in the volcano plots, and the correlation of some representative DEGs with GNG7 was presented in heatmaps. Gene Set Enrichment Analysis (GSEA) is a computational method for determining whether a defined set of genes shows statistically significant differences between two states. In the study, GSEA was performed by using the clusterProfiler R package (version 3.14.3) with c2 (c2.all.v7.0. entrez.gmt) from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005; Yu et al., 2012). Each analysis procedure was repeated 1000 times. The function or pathway termed with adjusted *p*-value <0.05 and false discovery rate (FDR) < 0.25 was considered statistically significant enrichment.

Tumor immune infiltration analysis, protein-protein interaction network analyses and the screening of hub genes

ESTIMATE algorithm was used to calculate the immune scores using the “estimate” R package (Yoshihara et al., 2013). Single-sample gene set enrichment analysis (ssGSEA) algorithm was used to assess the relative enrichment of the tumor tissue-infiltrating immune cells in LUAD (Hänzelmann et al., 2013). Based on an immune dataset for the 24 types of immunocytes, the relative enrichment score of every immunocyte was quantified from the gene expression profiles of each tumor sample (Bindea et al., 2013). In addition, we analyzed the differences in the enrichment of these 24 immune cells between the high and low GNG7 expression groups using ssGSEA. The Stat R package (Version 3.6.3) was employed to search for the genes related to GNG7 in LUAD. The correlation results were analyzed by the Spearman coefficient and the cut-off thresholds were |R| >0.4 and *p*-value <0.05. The list of immune-related genes was obtained from the ImmPort database. Genes associated with GNG7 were intersected with IRGs to obtain IRGs associated with GNG7. To further understand the interactions between IRGs associated with GNG7, we constructed a Protein-Protein Interaction (PPI) using the Search Tool for the Retrieval of Interacting Genes (STRING) (https://string-db.org/). An interaction with a combined score >0.4 was considered statistically significant. Cytoscape (Version 3.7.2) was used to visualize the network, while the cytoHubba plugin was used to rank genes within this network based upon their degree centrality values. Hub genes were considered to be those with the top 10 highest degree values. ClusterProfiler R package was employed to perform Gene Ontology (GO) function enrichment analyses. Furthermore,

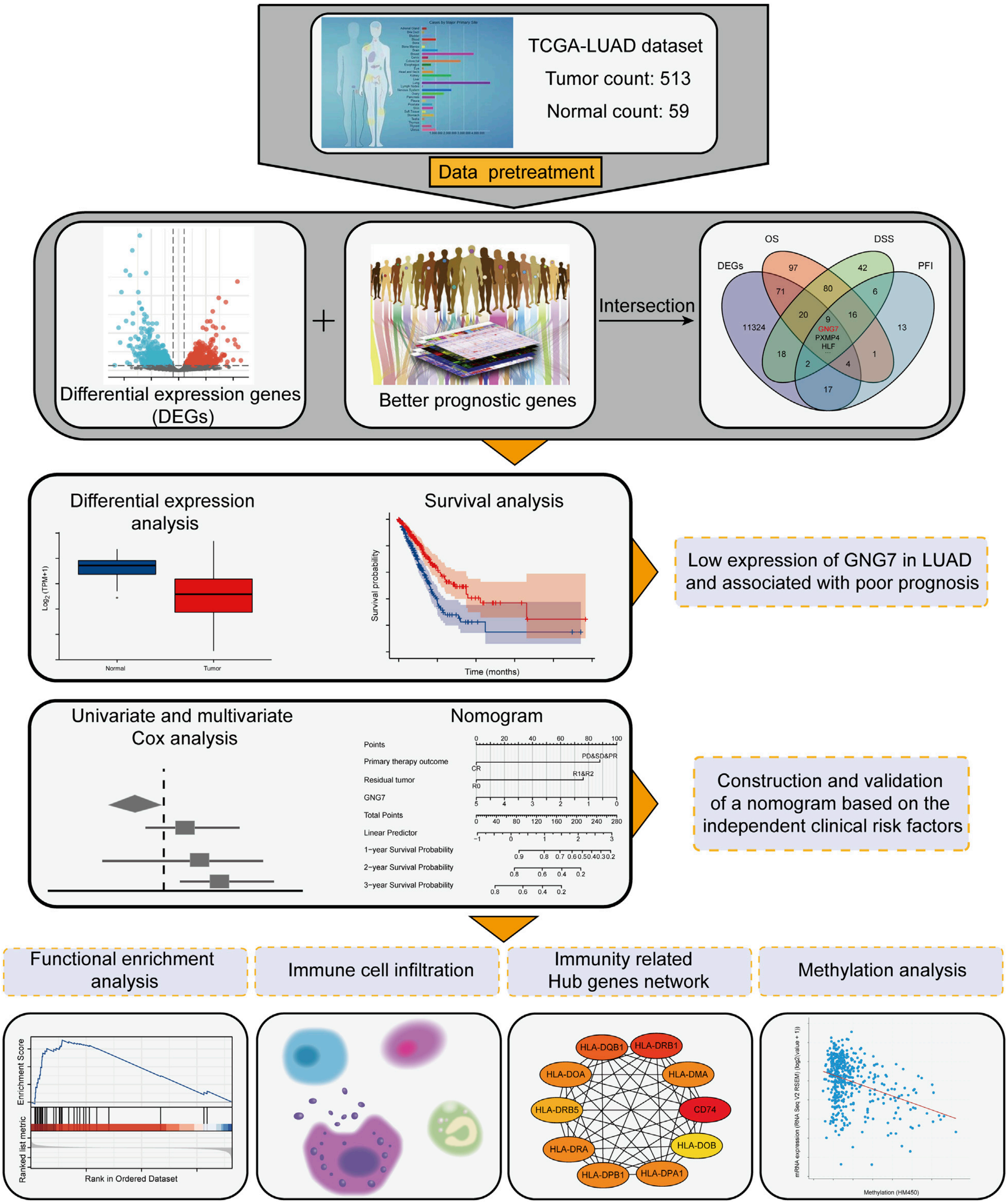


FIGURE 1
Workflow for screening of key genes and downstream analysis.

We used the GEPIA2 database (http://gepia2.cancer-pku.cn/#index) to analyze the correlation between GNG7 and B cell infiltration in LUAD tissues and normal tissues (Tang et al., 2019). Tumor Immune Estimation Resource 2.0 (TIMER2.0) database (http://timer.cistrome.org/) and the Kaplan-Meier plotter database (http://kmplot.com/analysis/) were used for the prognostic analysis of LUAD patients with different GNG7 expression and B cell infiltration (Li T. et al., 2020; Lánczky and Győrffy, 2021). We also used the TIMER database to analyze the correlation between GNG7 and immune cell markers in LUAD (Li et al., 2017).

DNA methylation analysis

To explore the possible mechanism of decreased expression of GNG7 in LUAD, we performed a differential methylation analysis of GNG7 between the normal and LUAD tissues using the UALCAN database. The cBio Cancer Genomics Portal (cBioPortal) (https://www.cbioportal.org), developed based on the TCGA database, was used to perform a correlation analysis between GNG7 mRNA expression and its methylation levels (Cerami et al., 2012; Gao et al., 2013). DNA methylation of GNG7 at CpG sites and the prognostic value of these CpG sites in LUAD were analyzed by MethSurv (https://biit.cs.ut.ee/methsurv/) (Modhukur et al., 2018).

Statistical analysis

All statistical analyses were conducted using R (Version 3.6.3). A part of the figures was plotted using the ggplot2 R package (Version 3.3.3). Dunn's test, Kruskal–Wallis test, and logistic regression were used to analyze the clinicopathological features of GNG7 in LUAD. Kaplan-Meier survival analysis, univariate and multivariate Cox regression analysis were performed for prognostic analysis. In all analyses, the *p*-value<0.05 was considered statistically significant. The specific datasets, R packages, software and databases used in each part of this study are detailed in Supplementary Table S4.

Results

G protein subunit gamma 7 is found to be one of the key regulators of lung adenocarcinoma tightly related to the prognosis through large-scale screening

To find key regulators of LUAD, we conducted screening work based on differential expression analysis and prognostic analysis, and performed a series of analytical work based on the target gene. The workflow was shown in Figure 1 we applied the DESeq2 R package to screen for DEGs in LUAD based on TCGA-LUAD datasets. The results showed that there were 7741 up-regulated and 3783 down-regulated genes among the screened 11,524 DEGs (Figure 2A). Combined with further prognostic analysis, we found that among the DEGs, nine genes (HSD17B6, PXMP4, HLF, ADGRD1, CYP17A1, ESYT3, FCAMR, C11orf16, GNG7) were significantly and positively associated with LUAD prognostic indicators including OS, DSS and PFI (Figure 2B). Of note, although GNG7 has been reported to be differentially expressed in a variety of tumors, its roles in the initiation and progression of LUAD remain unclear. In the present study, we focused on GNG7 to explore the underlying mechanism and clinical significance in LUAD.

G protein subunit gamma 7 expression is downregulated in lung adenocarcinoma

To elucidate the expression pattern of GNG7 in cancers, we first evaluated the expression of GNG7 in 33 types of cancers by a systematic analysis based on the TCGA databases. The results showed that GNG7 expression was significantly down-regulated in 17 different tumors, including Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Lung adenocarcinoma (LUAD) while it was significantly up-regulated in Cholangiocarcinoma (CHOL), Liver hepatocellular carcinoma (LIHC) and Pheochromocytoma and Paraganglioma (PCPG) (Figure 2C). Then, the low expression of GNG7 in LUAD was further validated by using three GEO datasets (GSE32665, GSE32863, GSE43458) (Figures 2D–F). The paired analysis got the similar results (Figure 2G). In addition, the decreased protein level of GNG7 was also observed in LUAD using the UALCAN database (Figure 2H). Moreover, ROC curve analysis was employed to analyze the distinguishing efficacy of GNG7 between LUAD tissue and normal tissue. The area under the curve (AUC) of GNG7 is 0.871, suggesting that GNG7 may be an ideal biomarker to distinguish LUAD from normal tissue (Figure 2I). Together, these results indicated that GNG7 is lowly expressed in LUAD which may be a potential diagnostic marker for LUAD.

Association between clinicopathological characteristics and GNG7 expression in lung adenocarcinoma

To clarify the correlation between the expression of GNG7 and clinicopathological variables, we collected data from the TCGA database on 535 patients with LUAD. After data preprocessing, the relationship between gene expression profiles and clinicopathological characteristics of 513 LUAD patients was shown in the baseline data table (Supplementary

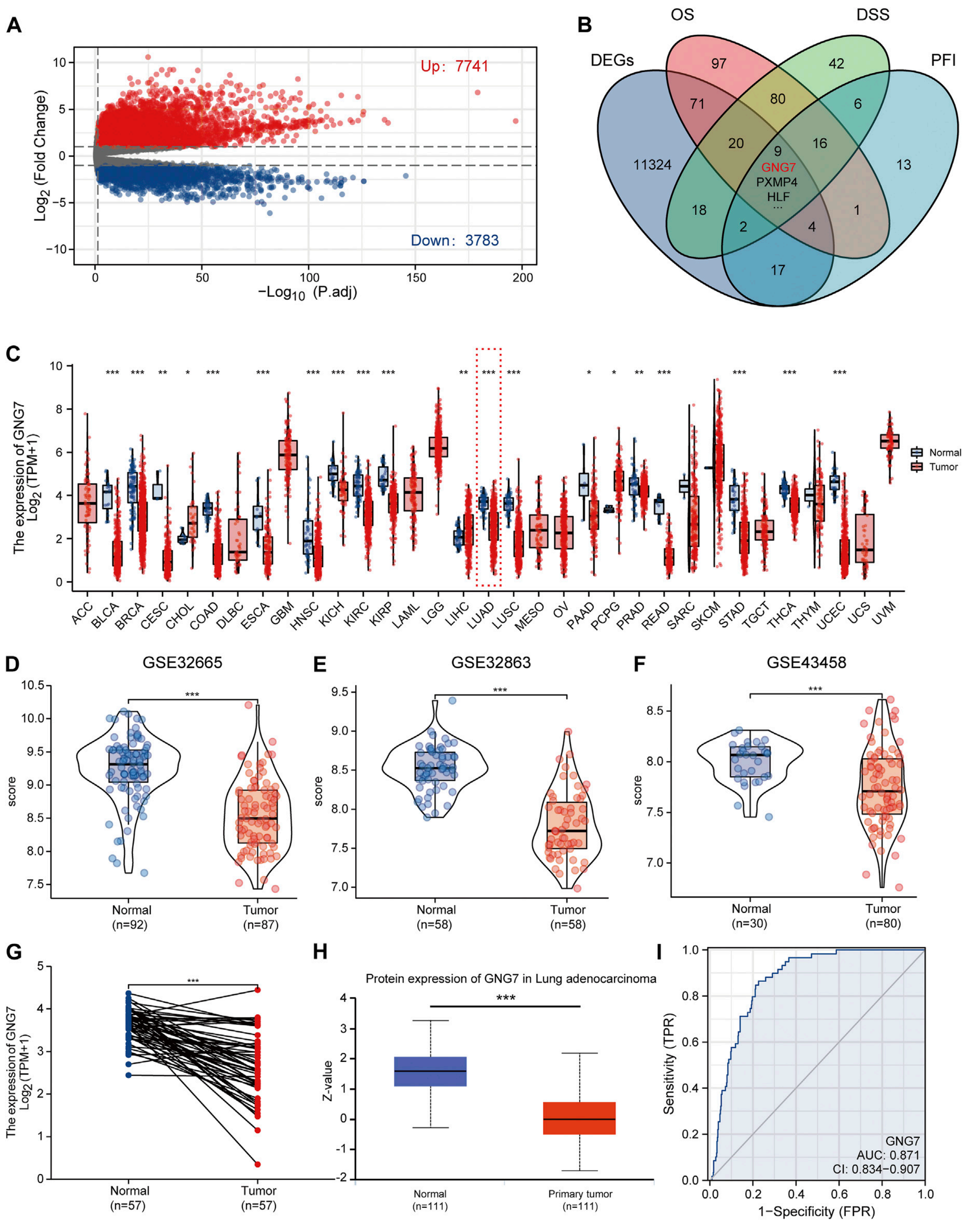


FIGURE 2
Screening of key regulators and identification of the differential expression of GNG7 in LUAD. **(A)** The volcanic map of the DEGs in LUAD. **(B)** A Venn diagram used to identify nine DEGs associated with LUAD prognostic indicators. **(C)** The GNG7 expression in different cancer from the TCGA database. **(D–F)** The GNG7 mRNA expression between LUAD and normal tissues based on data from GSE32665 **(D)**, GSE32863 **(E)** and GSE43458 **(F)** dataset. **(G)** The GNG7 mRNA expression between paired LUAD tumor tissues and adjacent normal tissues from the TCGA-LUAD dataset. **(H)** The GNG7 protein expression between LUAD and normal tissues from the UALCAN database. **(I)** A ROC curve to test the efficiency of GNG7 to identify LUAD from normal lung tissue. **p* < 0.05, ***p* < 0.01, ****p* < 0.001.

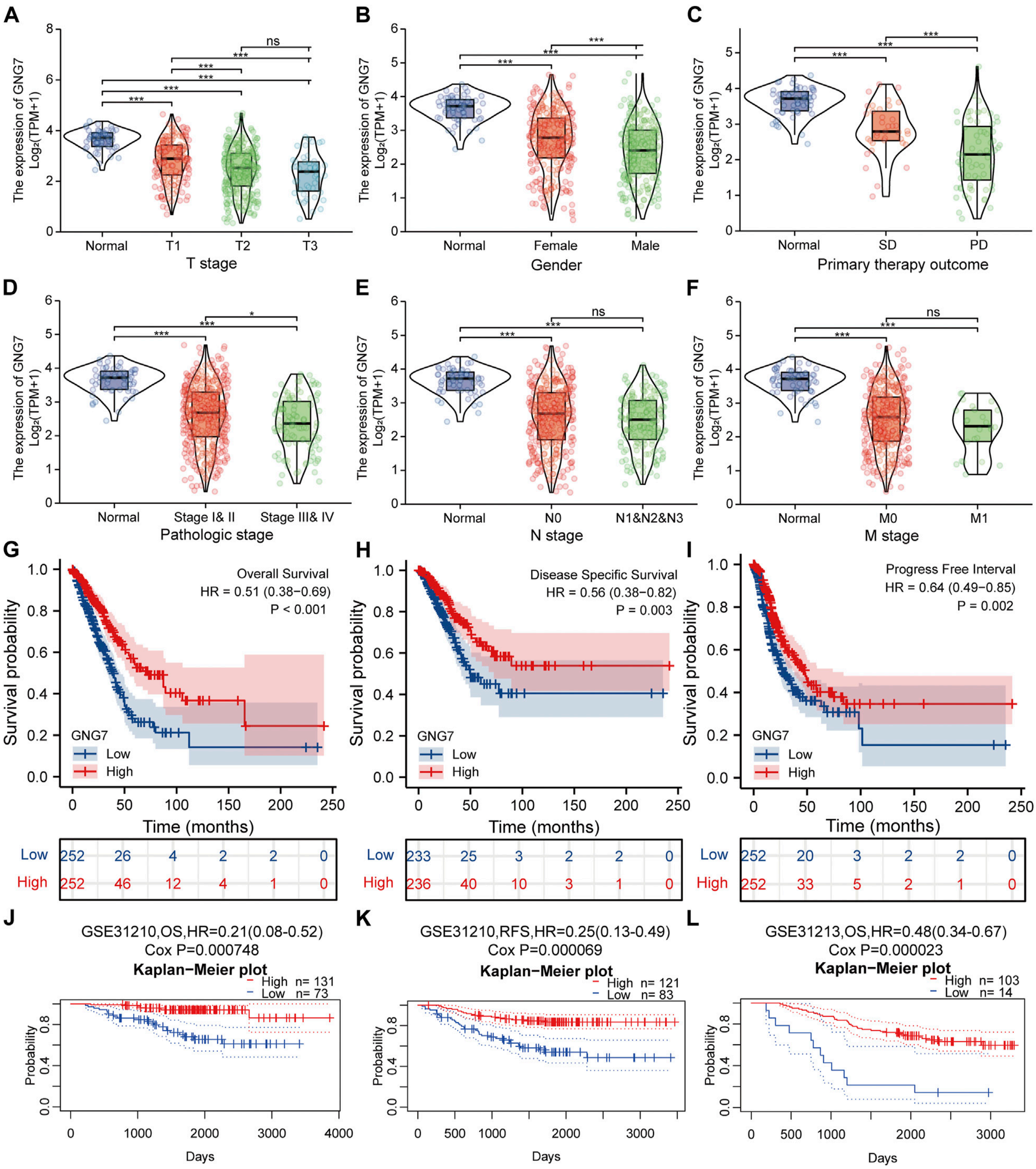


FIGURE 3
Correlation between GNG7 expression and clinicopathological features as well as the prognostic value of GNG7 in LUAD. **(A)** T stage. **(B)** Gender. **(C)** Primary therapy outcome. **(D)** Pathologic stage. **(E)** N stage. **(F)** M stage. ns, no significant difference, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. **(G–I)** Survival curves of OS **(G)**, DSS **(H)**, and PFI **(I)** between GNG7-high and GNG7-low expression groups from the TCGA-LUAD dataset. **(J–L)** Kaplan-Meier survival curve analysis of OS **(J)** and RFS **(K)** in a LUAD cohort (GSE31210) as well as OS **(L)** in a LUAD cohort (GSE31213) from the PrognoScan database.

TABLE 1 GNG7 expression association with clinical pathological characteristics (logistic regression).

Characteristics	Total (N)	Odds Ratio (OR)	*p* value
T stage (T2&T3&T4 vs. T1)	510	0.403 (0.274–0.589)	***<0.001
N stage (N1&N2&N3 vs. N0)	501	0.700 (0.482–1.014)	0.060
M stage (M1 vs. M0)	369	0.493 (0.197–1.140)	0.110
Pathologic stage (Stage II&Stage III&Stage IV vs. Stage I)	505	0.538 (0.377–0.766)	***<0.001
Primary therapy outcome (PD vs. SD)	105	0.311 (0.133–0.707)	**0.006
Gender (Male vs. Female)	513	0.502 (0.352–0.713)	***<0.001
Race (Black or African American&White vs. Asian)	446	2.701 (0.576–19.003)	0.238
Age (>65 vs. <=65)	494	1.050 (0.738–1.496)	0.785
Residual tumor (R1&R2 vs. R0)	361	0.565 (0.191–1.519)	0.271
Anatomic neoplasm subdivision (Right vs. Left)	498	0.957 (0.669–1.371)	0.812
Anatomic neoplasm subdivision2 (Peripheral Lung vs. Central Lung)	189	1.022 (0.555–1.894)	0.943
number_pack_years_smoked (>=40 vs. <40)	351	0.863 (0.567–1.312)	0.491
Smoker (Yes vs. No)	499	0.629 (0.378–1.035)	0.070

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table S1). The results showed that low expression of GNG7 was positively associated with high T stage, Gender (male sex), poor primary therapy outcome and high pathologic stage of LUAD, while there were no significant associations between GNG7 expression and the other clinical factors such as N stage and M stage (Figures 3A–F). In line with these findings, the logistics regression analysis also revealed that GNG7 expression was significantly associated with T stage (OR = 0.403, 95% CI: 0.274-0.589, $p < 0.001$), Pathologic stage (OR = 0.538, 95% CI: 0.377-0.766, $p < 0.001$), Primary therapy outcome (OR = 0.311, 95% CI: 0.133-0.707, $p = 0.006$) and Gender (OR = 0.502, 95% CI: 0.352-0.713, $p < 0.001$) (Table 1).

Significance of G protein subunit gamma 7 in clinical prognosis of lung adenocarcinoma and clinical subgroup analysis

We utilized data from the TCGA database to investigate the prognostic significance of GNG7 in LUAD. Kaplan-Meier survival analysis based on the TCGA-LUAD dataset revealed that low expression of GNG7 was associated with poor OS (HR = 0.51, 95% CI: 0.38-0.69, $p < 0.001$), DSS (HR = 0.56, 95% CI: 0.38-0.82, $p = 0.003$) and PFI (HR = 0.64, 95% CI: 0.49-0.85, $p = 0.002$) (Figures 3G–I). To further validate the prognostic value of GNG7 in LUAD, we utilized the PrognoScan database for further study. We included two of the GSE datasets (GSE31210 and GSE13213) in our analysis, where low GNG7 expression was significantly associated with the poorer prognosis (OS, HR = 0.21, 95% CI: 0,08-0.52, Cox $p = 0.000748$; RFS, HR = 0.25, 95% CI: 0,13-0.49, Cox $p = 0.000069$ in the GSE31210 dataset; OS, HR = 0.48, 95% CI: 0,34-0.67, Cox $p = 0.000023$ in the GSE13213 dataset) (Figures 3J–L).

Moreover, the univariate Cox regression analysis model showed that GNG7 expression level was significantly associated with OS (HR: 0.702; 95% CI: 0.599-0.822; $p < 0.001$) similar to T stage, N stage, M stage and Pathologic stage as well Primary therapy outcome and Residual tumor. Meanwhile, the multivariate Cox regression analysis also revealed that low expression of GNG7, similar to Primary therapy outcome and Residual tumor, was an independent risk factor for the prognosis of LUAD patients (Table 2). Collectively, these results suggest that low expression of GNG7 independently predicts poor prognosis for patients with LUAD.

Given that multivariate Cox regression analysis identified low expression of GNG7 as an independent risk factor, we investigated the potential prognostic value of GNG7 in LUAD patients with different clinical subgroups. As shown in Figures 4A–C, low expression of GNG7 was associated with poor prognosis in stage N0, including OS (HR = 0.42, 95% CI: 0.27-0.67, $p < 0.001$), DSS (HR = 0.38, 95% CI: 0.21-0.69, $p = 0.001$) and PFI (HR = 0.57, 95% CI: 0.40-0.83, $p = 0.003$). However, there was no statistically significant correlation between GNG7 expression and prognosis in the N1&N2&N3 stage ($p > 0.05$) (Supplementary Figures S1A–C). In addition, low GNG7 expression was significantly associated with poor prognosis in LUAD patients in M0 stage, including OS (HR = 0.48, 95% CI: 0.33-0.68, $p < 0.001$), DSS (HR = 0.59, 95% CI: 0.37-0.94, $p = 0.026$), PFI (HR = 0.66, 95% CI: 0.47-0.92, $p = 0.015$) (Figures 4D–F). Nevertheless, no significant association was shown between GNG7 expression and prognosis in LUAD patients in the M1 stage (Supplementary Figures S1D–F). These results suggest that low expression of

TABLE 2 Univariate analysis and multivariate analysis of the correlation between clinicopathological characteristics and OS in LUAD.

Characteristics	Total (N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	*p* value	Hazard ratio (95% CI)	*p* value
T stage (T2&T3&T4 vs. T1)	501	1.668 (1.184–2.349)	**0.003	1.116 (0.629–1.978)	0.708
N stage (N1&N2&N3 vs. N0)	492	2.606 (1.939–3.503)	***<0.001	1.689 (0.758–3.764)	0.200
M stage (M1 vs. M0)	360	2.111 (1.232–3.616)	**0.007	1.698 (0.674–4.280)	0.262
Pathologic stage (Stage II&Stage III&Stage IV vs. Stage I)	496	2.975 (2.188–4.045)	***<0.001	1.109 (0.471–2.610)	0.812
Primary therapy outcome (PD&SD&PR vs. CR)	419	2.818 (2.004–3.963)	***<0.001	3.662 (2.217–6.049)	***<0.001
Gender (Male vs. Female)	504	1.060 (0.792–1.418)	0.694		
Race (White vs. Asian&Black or African American)	446	1.422 (0.869–2.327)	0.162		
Age (>65 vs. <=65)	494	1.228 (0.915–1.649)	0.171		
Residual tumor (R1&R2 vs. R0)	352	3.973 (2.217–7.120)	***<0.001	3.670 (1.503–8.964)	**0.004
Anatomic neoplasm subdivision (Right vs. Left)	490	1.024 (0.758–1.383)	0.878		
Anatomic neoplasm subdivision2 (Peripheral Lung vs. Central Lung)	182	0.913 (0.570–1.463)	0.706		
Number pack years smoked (>=40 vs. <40)	345	1.038 (0.723–1.490)	0.840		
Smoker (Yes vs. No)	490	0.887 (0.587–1.339)	0.568		
GNG7	504	0.702 (0.599–0.822)	***<0.001	0.727 (0.561-0.943)	*0.016

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

GNG7 is positively associated with the poor prognosis of LUAD patients without lymph node invasion and distal metastasis.

Construction and validation of a nomogram based on the independent clinical risk factors

To provide a quantitative approach to predicting the prognosis of LUAD patients, we constructed a prognostic nomogram to predict individual survival probability based on the expression levels of GNG7 and other independent clinical risk factors (Figure 4G). The calibration curve of the nomogram showed that the established lines of 1-, 2-, and 3-y survival highly matched the ideal line (the 45-degree line) (Figure 4H). In addition, the C-index of the prediction model reached 0.690 (0.659–0.720), indicating that the model had a reliable potential to predict the OS of LUAD patients. In addition, ROC curve analysis based on the three time points of 1-, 2-, and 3-Year showed that the Area under the curve (AUC) of this prediction model was higher than the AUC of the two-by-two model consisting of independent prognostic factors screened from multivariate Cox regression, indicating the superiority of the model (Supplementary Figures S2A–D). On the basis of the median risk score, patients were divided into a high-risk score group and a low-risk score group. Survival curve analysis revealed that the high-risk group had a significantly poorer prognosis compared to the low-risk group (HR = 2.59, 95% CI: 2.59 (1.72–3.89), $p < 0.001$) (Supplementary Figure S2E). Additionally, the risk curve indicated that the high-risk score group had higher mortality and worse prognosis than the low-risk score group (Supplementary Figure S2F).

Functional enrichment and pathway analysis of G protein subunit gamma 7-associated differentially expressed genes in lung adenocarcinoma

To investigate the biological functions and signaling pathways associated with GNG7, we examined the DEGs between GNG7-high and GNG7-low patients which were stratified based on the median GNG7 expression. Resultantly, 1403 mRNAs (492 upregulated and 911 downregulated), 962 lncRNAs (256 upregulated and 706 downregulated), and 21miRNAs (18 upregulated and 3 downregulated) were differently expressed in GNG7-high patients compared to GNG7-low ones (Figure 5A, Supplementary Figures S3A,C). Relative expression values of some representative DEGs between the two cohorts were shown in the form of heatmaps (Figure 5B, Supplementary Figures S3B,D). Strikingly, pathway enrichment analysis showed that the DEGs were most strongly enriched in the B cell receptor signaling pathway, T cell receptor signaling pathway and HIV infection allograft rejection which are highly

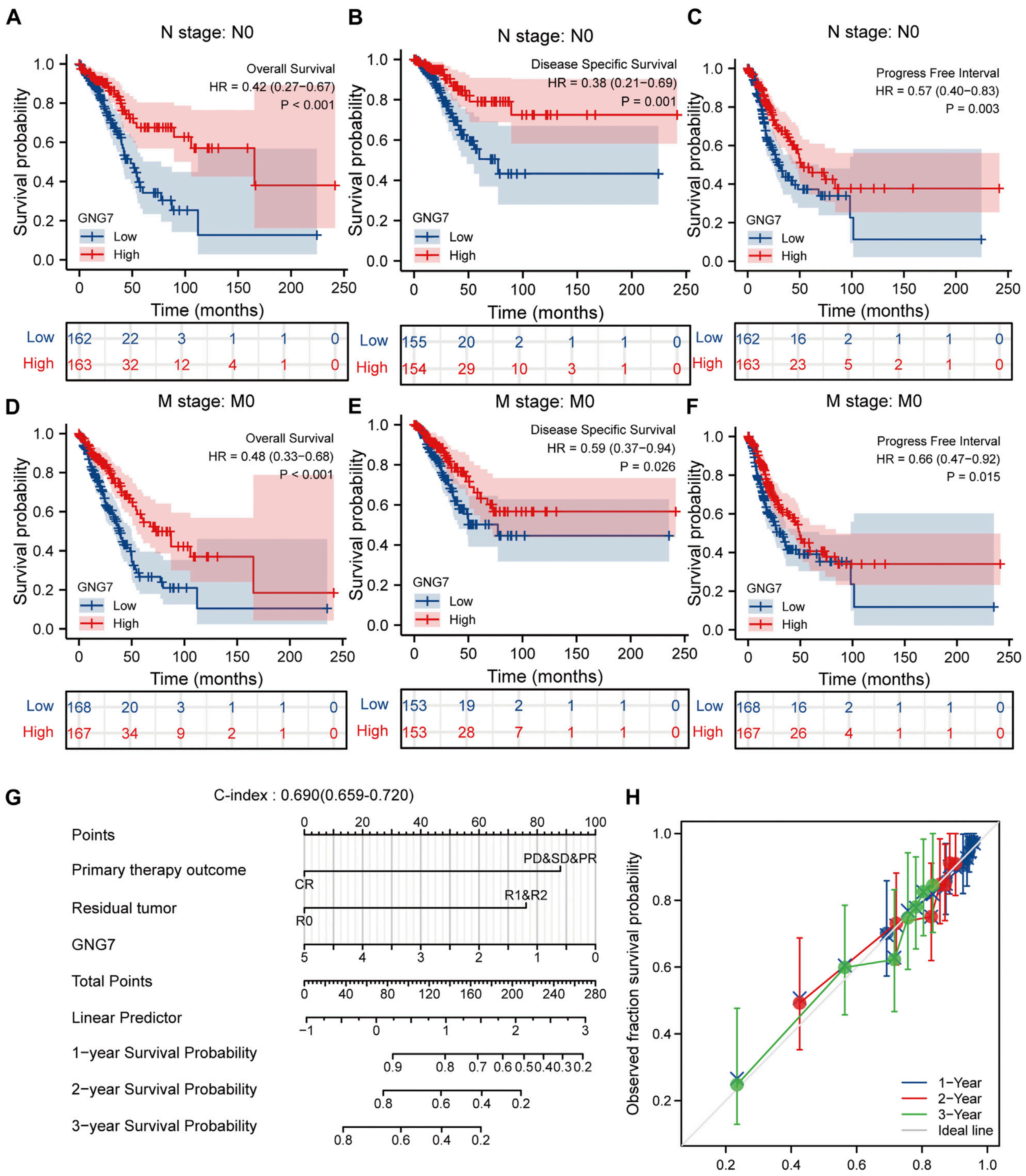


FIGURE 4
Subgroup analysis and the construction of a nomogram based on GNG7 expression. **(A–C)** The Kaplan-Meier curves of OS **(A)**, DSS **(B)**, and PFI **(C)** between GNG7-high and -low expression patients with LUAD in N0 stage. **(D–F)** The Kaplan-Meier curves of OS **(D)**, DSS **(E)**, and PFI **(F)** between GNG7-high and -low expression patients with LUAD in M0 stage. **(G)** A nomogram that integrates GNG7 and other independent prognostic factors in LUAD from TCGA data. **(H)** The calibration curve of the nomogram. OS, overall survival; DSS, disease specific survival; PFI, progress free interval; LUAD, lung adenocarcinoma.

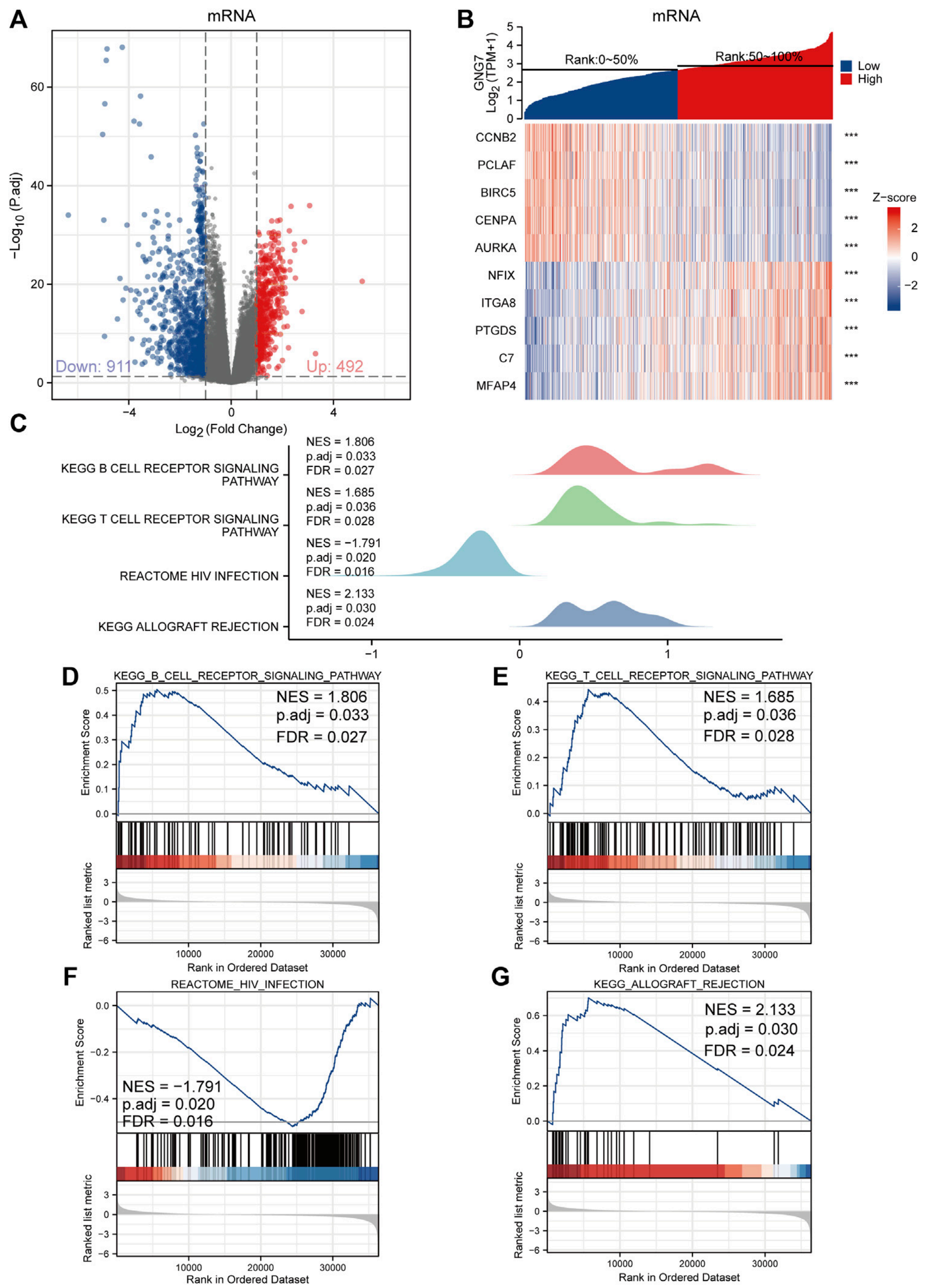
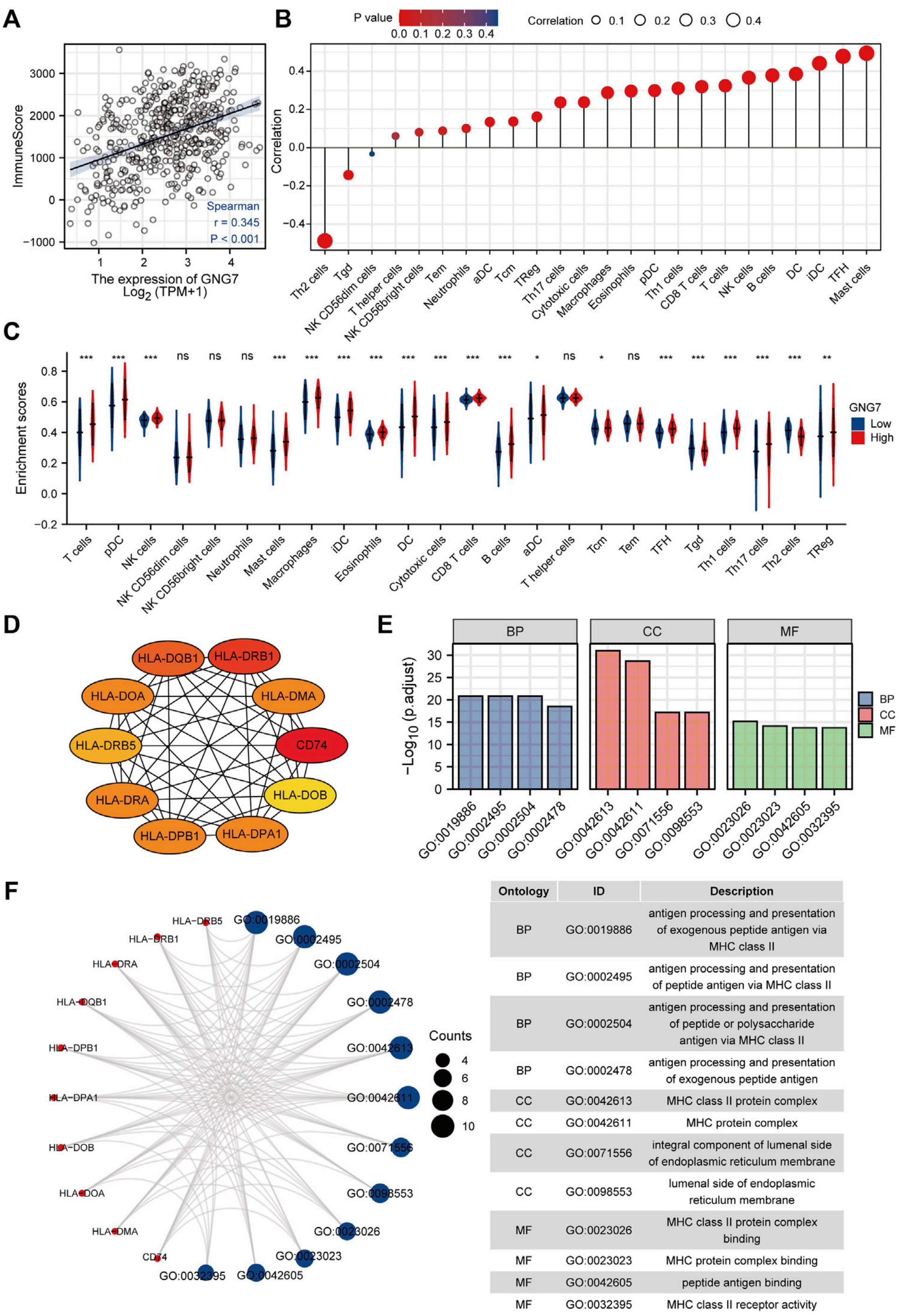


FIGURE 5
Functional annotation of differentially expressed genes (DEGs) regulated by GNG7 in LUAD. **(A,B)** Based on the median GNG7 expression level, LUAD patients from the TCGA-LUAD dataset were stratified into GNG7-high and GNG7-low groups. Expression profiles of mRNAs in two groups are presented by volcano plots **(A)** and heatmaps **(B)**. **(C)** Pathway enrichment plots from GSEA. **(D–G)** The B cell receptor signaling pathway **(D)**, T cell receptor signaling pathway **(E)**, HIV infection **(F)**, and allograft rejection **(G)** were positively correlated to GNG7 expression. TCGA, the cancer genome atlas; LUAD, lung adenocarcinoma; NES, normalized enrichment score; FDR, false discovery rate.



Ontology	ID	Description
BP	GO:0019886	antigen processing and presentation of exogenous peptide antigen via MHC class II
BP	GO:0002495	antigen processing and presentation of peptide antigen via MHC class II
BP	GO:0002504	antigen processing and presentation of peptide or polysaccharide antigen via MHC class II
BP	GO:0002478	antigen processing and presentation of exogenous peptide antigen
CC	GO:0042613	MHC class II protein complex
CC	GO:0042611	MHC protein complex
CC	GO:0071556	integral component of lumenal side of endoplasmic reticulum membrane
CC	GO:0098553	lumenal side of endoplasmic reticulum membrane
MF	GO:0023026	MHC class II protein complex binding
MF	GO:0023023	MHC protein complex binding
MF	GO:0042605	peptide antigen binding
MF	GO:0032395	MHC class II receptor activity

FIGURE 6
Correlation of immune cell infiltration and GNG7 expression in LUAD patients. **(A)** Relationship between immune scores and GNG7 expression levels in LUAD. **(B)** Relationships between infiltration levels of 24 immune cell types and GNG7 expression profiles by Spearman's analysis. **(C)** Comparison of the immune infiltration level of 24 immune cell types between GNG7-high and GNG7-low groups. LUAD, lung adenocarcinoma; DCs, dendritic cells; aDCs, activated DCs; iDCs, immature DCs; pDCs, plasmacytoid DCs; Th, T helper cells; Th1, type 1 Th cells; Th2, type 2 Th cells; Th17, type 17 Th cells; Treg, regulatory T cells; Tgd, T gamma delta; Tcm, T central memory; Tem, T effector memory; Tfh, T follicular helper; NK, natural killer. **(D)** The top10 hub genes calculated with the MCC algorithm by cytoHubb. **(E)** The top 4 enriched GO terms of BP, CC and MF categories of the GO enrichment analysis. **(F)** The network including the hub genes and the enriched GO terms. BP, biological process; CC, cellular component; MF, molecular function.

related to the cellular immune response (Figures 5C–G). These data suggested that GNG7 may play an important role in regulating the tumor immune microenvironment of LUAD.

Correlation analysis between the expression of G protein subunit gamma 7 and immune cell infiltration in lung adenocarcinoma

As reported, tumor-associated immune cell infiltration has a close relationship with tumor development and the prognosis of patients. Then, we utilized the ESTIMATE algorithm to assess the correlation between GNG7 and the abundance of immune cell infiltration in LUAD. The results revealed that GNG7 expression was positively correlated with the abundance of immune cell infiltration in LUAD (Figure 6A). Specifically, further Spearman correlation analysis showed that among 24 immune cell subpopulations, GNG7 expression was positively correlated with most immune cell subsets, including Mast cell, DC, B cells, and CD8$^+$ T cells, but negatively correlated with Th2 and Tgd cells (Figure 6B). Consistently, the ssGSEA analysis demonstrated that the infiltration levels of most of the immune cell subsets such as Mast cells, pDCs, B cells, NK cells and CD8$^+$ T cells were remarkably increased in LUAD patients with GNG7 high expression compared to those with GNG7 low expression (Figure 6C). In keeping with this finding, GNG7 was significantly correlated with most immune markers of different immune cells, including CD8$^+$T cell, B cell, Neutrophils, and Dendritic cells (Supplementary Table S2). Moreover, by correlation analysis, we identified immune-related genes (IRG) co-expressed with GNG7 and constructed a PPI network (Supplementary Figure S4A). We screened the top10 of hub genes and showed the correlation between these 10 genes and GNG7 in the form of scatter plots (Figure 6D, Supplementary Figures S4B–K). Additionally, we performed GO enrichment analysis to investigate the possible involvement of GNG7 in the immune response. The terms identified in the BP category showed that aberrantly expressed GNG7 was associated with antigen processing and presentation *via* MHC class II(MHCII), while in the CC category, the hub genes were significantly enriched in the MHCII protein complex and endoplasmic reticulum-related terms. Furthermore, the MF category revealed significant enrichment in GO terms related to the MHCII protein complex binding, MHCII receptor activity and peptide antigen binding, etc (Figures 6E,F). Together, these results suggest that GNG7 may contribute to the remodeling of the immune microenvironment in LUAD through promoting the infiltration of a variety of tumor-associated immune cells and influencing antigen presentation.

G protein subunit gamma 7 high expression with high B cell infiltration predicts a better prognosis of lung adenocarcinoma patients

Of the infiltrated immune cells increased in LUAD with GNG7 high expression, B cell infiltration attracts our attention as the relatively less knowledge of this cell type in tumor immunotherapy currently. Our results showed that GNG7 was closely associated with the level of B cell infiltration in LUAD (Supplementary Figure S5A). Specifically, the level of B cell infiltration was significantly elevated in the GNG7 high expression group compared with that in the GNG7 low expression group (Supplementary Figure S5B). In addition, we respectively investigated the correlation of GNG7 with B cells in the tumor and normal tissues. Strikingly, GNG7 showed a strong positive correlation with the B cell marker genes CD19 and CD79A in LUAD tissues (Figure 7A). In contrast, the correlation of GNG7 with the B cell markers CD19 and CD79A did not reach statistical significance in normal tissues (Figure 7B). These results suggest that GNG7 expression may promote B cell infiltration in the context of LUAD. Meanwhile, through KM plot database analysis, we found that patients with high GNG7 expression tended to predict a better prognosis in the B cell enriched group but not in the B cell decreased group (Figures 7C,D). Such finding was further corroborated by the analysis using the TIMER2.0 database, implying that high GNG7 expression corresponded to a better prognosis for LUAD patients in the context of enriched B cell infiltration. Furthermore, we found that high infiltration levels of B cells in the presence of consistent levels of GNG7 expression corresponded to a good prognosis in patients with LUAD (Figure 7E). Taken together, it is reasonable to suggest that GNG7 may have improved patient prognosis by promoting B cell infiltration.

G protein subunit gamma 7 dysregulation is associated with aberrant DNA methylation

Considering the importance of DNA methylation in regulating gene expression, we tested whether aberrant DNA methylation occurs in *GNG7* gene in LUAD. By analyzing the data from the UALCAN database, we found that the methylation level of GNG7 was significantly higher in the tumor group compared to that in the normal group (Figure 8A). Next, the correlation analysis based on the cBioPortal database showed that GNG7 expression was significantly negatively correlated with methylation (Figure 8B). To further investigate the methylation of GNG7 in LUAD, we analyzed the methylation levels of different CpG sites of GNG7 in LUAD patients using the MethSurv database and presented them in the form of heat maps (Figure 8C). The results revealed that several CpG sites of

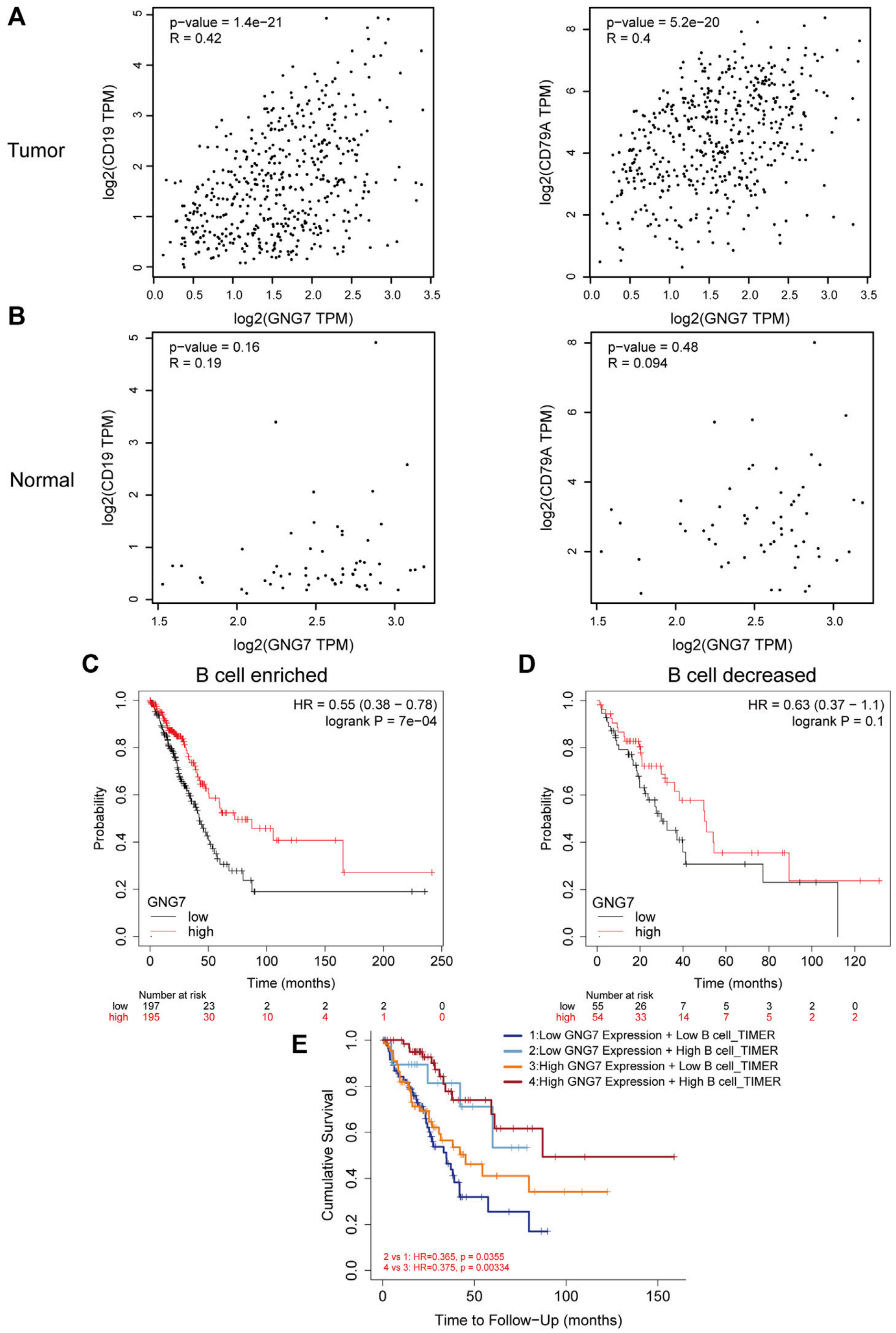


FIGURE 7
Correlation between GNG7 and B cell immune infiltration and prognostic analysis in LUAD. **(A)** The correlation between the expression of GNG7 and B cell markers CD19 (left) as well CD79A (right) in LUAD tissues. **(B)** The correlation between the expression of GNG7 and B cell markers CD19 (left) as well CD79A (right) in normal lung tissues. **(C)** Kaplan-Meier survival curves of OS in LUAD based on GNG7 expression in the enriched B cells groups. **(D)** Kaplan-Meier survival curves of OS based on GNG7 expression in the decreased B cells groups. **(E)** Kaplan-Meier survival curves of OS in LUAD based on B cell infiltration level in GNG7-high and -low expression patients. LUAD, lung adenocarcinoma; OS, overall survival.

GNG7 exhibited high methylation in LUAD patient samples, including cg19477361, cg21462934, and cg27181295. Prognostic analysis showed that the above CpG sites with highly methylated levels were associated with poor prognosis in LUAD (Figures 8D–F). These results suggest that the low expression of GNG7 in LUAD may be partly due to the methylation modification of the

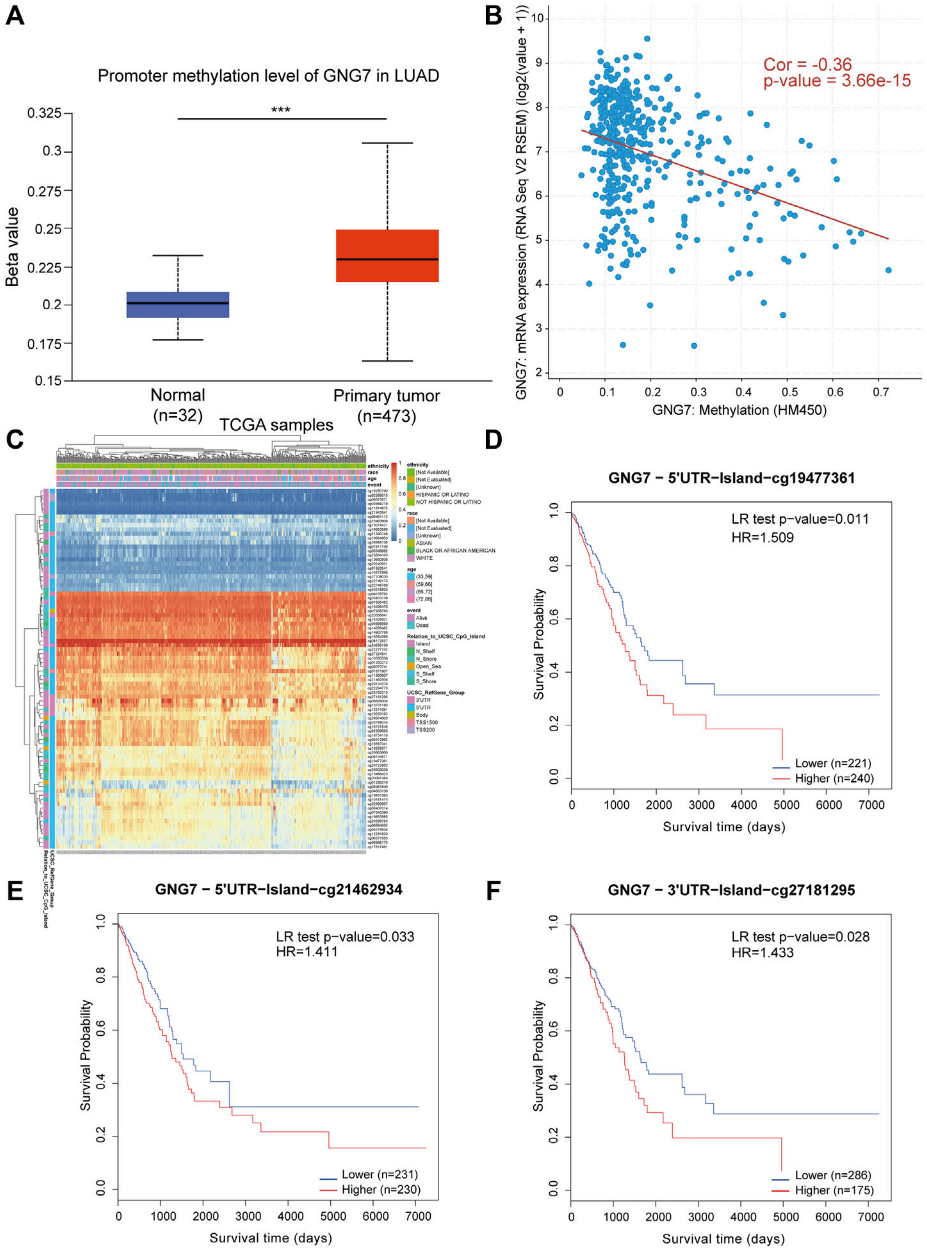


FIGURE 8
DNA methylation levels of GNG7 and its prognostic value in LUAD. **(A)** The promoter methylation level of GNG7 in normal tissues and primary LUAD tissues by the UALCAN database. **(B)** The correlation between GNG7 methylation and its expression level. **(C)** The heatmap of DNA methylation at CpG sites in the GNG7 gene by the MethSurv database. **(D–F)** Kaplan-Meier survival curves of OS based on methylation at GpG sites of cg19477361 **(D)**, cg21462934 **(E)** and cg27181295 **(F)** in LUAD.

abovementioned CpG sites and plays a key role in tumor progression.

Discussion

As one of the most common malignancies worldwide, the prognosis of LUAD patients remains very gloomy due to the lack of effective biomarkers for early diagnosis and effective treatment for advanced patients. Thus, intense research has been focused on deciphering the pathogenesis and searching for effective diagnostic and therapeutic approaches as well as prognostic markers to improve the prognosis of patients with LUAD in the last several years. Indeed, a growing number of potential biomarkers for LUAD have been identified, such as PPP1R14D, lncRNA-Ac068228, IFITM1, and so on (Koh et al., 2019; Jiang et al., 2022; Tian Y.

et al., 2022). However, most of these biomarkers are associated with the increase in cell numbers resulting from cell division (cell proliferation), programmed cell death (apoptosis), and tumor angiogenesis, while few with tumor immune microenvironment. Accumulating evidence has shown that not only the characteristics of tumor cells but also the tumor microenvironment, especially Tumor infiltrating immune cells (TIICs), plays critical roles in the tumorigenesis and progression of LUAD (Hinshaw and Shevde, 2019). In recent years, with the further understanding of the mechanism of tumor immune infiltration, tumor immunotherapies, such as the immune checkpoint inhibitors (ICIs), have had a revolutionary impact on the treatment of LUAD (Bagchi et al., 2021). However, only a small percentage of patients achieved a durable immune response after treatment. The mechanisms of LUAD development are far from being elucidated. It is of great necessity to further clarify the molecular basis of LUAD and explore contributing factors as well as sensitive diagnosis and prognosis biomarkers of immunotherapy response to improve patient outcomes (Wu and Shih, 2018; Singh et al., 2020).

In the current study, we performed comprehensive bioinformatics analyses to explore the potential key molecules involved in the development of LUAD. Through screening and identification, GNG7 was demonstrated to be lowly expressed in LUAD and had a good diagnostic performance. In addition, low expression of GNG7 was positively associated with the poor clinicopathological characteristics such as poor primary therapy outcome and high pathologic stage of LUAD, implying the tumor suppressive roles of GNG7 in LUAD.

As a subunit of heterotrimeric G protein, GNG7 has been reported to be tightly related to carcinogenesis. GNG7 is frequently downregulated in various cancers including pancreatic cancer, esophageal cancer and clear cell renal cell carcinoma (Shibata et al., 1998; Ohta et al., 2008; Xu et al., 2019). It is worth noting that GNG7 has been identified as one of the hub genes in an eight-gene prognostic signature model and a four-gene panel predicting overall survival for LUAD (Li C. et al., 2020; Ma et al., 2021). However, the role of GNG7 as an independent prognostic factor in LUAD has not been fully elucidated. Here, our study based on GNG7 is similar to, but more distinctive from, the newly identified biomarkers for LUAD in the latest literatures. In our study, we demonstrated GNG7 as an independent prognostic risk factor in LUAD (Supplementary Table S4) (Wan et al., 2021; Tian W. et al., 2022; Zhang et al., 2022; Zhou et al., 2022). Moreover, we constructed a nomogram combined with other clinical independent prognostic risk factors to predict the prognosis of LUAD patients reliably.

Recent studies have reported that GNG7 inhibited the progression of LUAD by inhibiting E2F1 and Hedgehog signaling, but the exact mechanism by which it regulates the development of LUAD is largely unknown (Zhao et al., 2021; Zheng et al., 2021). Our GSEA and GO enrichment analysis found that GNG7 may be involved in regulating the TME of LUAD and antigen processing and presentation *via* MHCII. Especially, high expression of GNG7 corresponded to increased infiltration level of several immune cells including B cells. Over the past decades, intense research has focused on the roles of T cells in immune regulation in the TME (Joyce and Fearon, 2015). More recently, there is increasing evidence supporting a critical role for B cells in tumor immunology (Bruno, 2020; Cabrita et al., 2020; Helmink et al., 2020). However, there is limited understanding of the biological contributors to the B cell infiltration in the TME. As the key immune cell in humoral immunity, B cells express a large number of MHCII molecules and are important antigen-presenting cells. In the present study, we found that GNG7 may be involved in regulating MHCII-mediated antigen processing and presentation. Furthermore, we found that GNG7 expression may promote B cell infiltration as evidenced by that low GNG7 expression was negatively correlated with B cell infiltration. Strikingly, the results of the prognostic analysis indicated patients with high GNG7 expression tended to predict a better prognosis in the context of enriched B cell infiltration and high infiltration levels of B cells regardless of GNG7-low or GNG7-high expression corresponded to a good prognosis in patients with LUAD. Our results indicated that GNG7 may exert its tumor suppressive roles in LUAD by promoting B cell infiltration and GNG7 expression together with B cell infiltration may be a powerful predictive signature for prognosis and immunotherapy response in LUAD, although the detailed function and mechanism need further in-depth investigation both *in vitro* and *in vivo*.

Finally, in this study, we further explored the mechanism of GNG7 low expression in LUAD. To our knowledge, other than miR-19b-3p which was reported to target GNG7 directly and significantly decrease the mRNA level of GNG7, little is known about the mechanism of GNG7 dysregulation in LUAD (Zhao et al., 2021). Given that DNA methylation of CpG islands is known to be a repressive mark of gene expression, we assessed the methylation level of GNG7 in LUAD (Kulis and Esteller, 2010). As expected, elevated methylation level of GNG7 was observed in tumor tissues which may be responsible for the low expression of GNG7 in LUAD. Interestingly, hypermethylation of GNG7 was associated with poor prognosis in patients with LUAD, which is consistent with the prognostic value of GNG7 mRNA expression. These results suggest the importance of DNA methylation in regulating GNG7 expression in LUAD.

In summary, this study revealed for the first time that GNG7 may be involved in regulating the immune microenvironment in LUAD and influence tumor development and patient prognosis at least partly by regulating the B cell infiltration. GNG7 may be not only a potential diagnostic biomarker for LUAD but also a promising predictive signature for prognosis and immunotherapy response for patients with LUAD. Nevertheless, as data from this study were mainly obtained from open databases, more LUAD patient samples are needed to confirm the clinical prognostic value of GNG7. Moreover, the effects of GNG7 on immune cell recruitment and infiltration as well as immunotherapy response are needed to be investigated deeply at the cellular and molecular levels and in future clinical trials.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: TCGA database: https://portal.gdc.cancer.gov/ and GEO database: https://www.ncbi.nlm.nih.gov/geo/.

Author contributions

HY and QW designed the study. QW performed the acquisition and analysis of data. QW wrote the manuscript. PZ, TM, BJ, and HY edited the manuscript. All authors read and approved the final manuscript.

Funding

This work was funded by National Natural Science Foundation of China (Nos 31870781) and Shandong Province Science and Technology Development Plan Foundation for Traditional Chinese Medicine (Nos 2017-047).

Acknowledgments

We would like to thank the National Natural Science Foundation of China (Nos 31870781) and the Shandong Provincial Science and Technology Development Program Fund for Traditional Chinese Medicine (Nos 2017-047) for their financial support, as well as TCGA and GEO databases for providing platforms and contributors to upload their meaningful datasets. It should be noted that this article has been shared as a preprint on the Research Square platform under the DOI number: https://doi.org/10.21203/rs.3.rs-1702007/v1.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.984575/full#supplementary-material

References

Bagchi, S., Yuan, R., and Engleman, E. G. (2021). Immune checkpoint inhibitors for the treatment of cancer: clinical impact and mechanisms of response and resistance. *Annu. Rev. Pathol.* 16, 223–249. doi:10.1146/annurev-pathol-042020-042741

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A. C., et al. (2013). Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* 39 (4), 782–795. doi:10.1016/j.immuni.2013.10.003

Bruno, T. C. (2020). New predictors for immunotherapy responses sharpen our view of the tumour microenvironment. *Nature* 577 (7791), 474–476. doi:10.1038/d41586-019-03943-0

Cabrita, R., Lauss, M., Sanna, A., Donia, M., Skaarup Larsen, M., Mitra, S., et al. (2020). Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* 577 (7791), 561–565. doi:10.1038/s41586-019-1914-8

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2 (5), 401–404. doi:10.1158/2159-8290.Cd-12-0095

Chandrashekar, D. S., Karthikeyan, S. K., Korla, P. K., Patel, H., Shovon, A. R., Athar, M., et al. (2022). UALCAN: an update to the integrated cancer data analysis platform. *Neoplasia* 25, 18–27. doi:10.1016/j.neo.2022.01.001

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6 (269), pl1. doi:10.1126/scisignal.2004088

Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinforma.* 14, 7. doi:10.1186/1471-2105-14-7

Hartmann, S., Szaumkessel, M., Salaverria, I., Simon, R., Sauter, G., Kiwerska, K., et al. (2012). Loss of protein expression and recurrent DNA hypermethylation of the GNG7 gene in squamous cell carcinoma of the head and neck. *J. Appl. Genet.* 53 (2), 167–174. doi:10.1007/s13353-011-0079-4

Helmink, B. A., Reddy, S. M., Gao, J., Zhang, S., Basar, R., Thakur, R., et al. (2020). B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* 577 (7791), 549–555. doi:10.1038/s41586-019-1922-8

Hinshaw, D. C., and Shevde, L. A. (2019). The tumor microenvironment innately modulates cancer progression. *Cancer Res.* 79 (18), 4557–4566. doi:10.1158/0008-5472.Can-18-3962

Jiang, X., Chen, M., Du, J., Bi, H., Guo, X., Yang, C., et al. (2022). LncRNA-AC068228.1 is a novel prognostic biomarker that promotes malignant phenotypes in lung adenocarcinoma. *Front. Oncol.* 12, 856655. doi:10.3389/fonc.2022.856655

Joyce, J. A., and Fearon, D. T. (2015). T cell exclusion, immune privilege, and the tumor microenvironment. *Science* 348 (6230), 74–80. doi:10.1126/science.aaa6204

Koh, Y. W., Han, J. H., Jeong, D., and Kim, C. J. (2019). Prognostic significance of IFITM1 expression and correlation with microvessel density and epithelial-mesenchymal transition signature in lung adenocarcinoma. *Pathol. Res. Pract.* 215 (7), 152444. doi:10.1016/j.prp.2019.152444

Kulis, M., and Esteller, M. (2010). DNA methylation and cancer. *Adv. Genet.* 70, 27–56. doi:10.1016/b978-0-12-380866-0.60002-2

Lánczky, A., and Győrffy, B. (2021). Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. *J. Med. Internet Res.* 23 (7), e27633. doi:10.2196/27633

Li, C., Long, Q., Zhang, D., Li, J., and Zhang, X. (2020a). Identification of a four-gene panel predicting overall survival for lung adenocarcinoma. *BMC Cancer* 20 (1), 1198. doi:10.1186/s12885-020-07657-9

Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77 (21), e108–e110. doi:10.1158/0008-5472.Can-17-0307

Li, T., Fu, J., Zeng, Z., Cohen, D., Li, J., Chen, Q., et al. (2020b). TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* 48 (W1), W509–w514. doi:10.1093/nar/gkaa407

Liu, J., Ji, X., Li, Z., Yang, X., Wang, W., and Zhang, X. (2016). G protein γ subunit 7 induces autophagy and inhibits cell division. *Oncotarget* 7 (17), 24832–24847. doi:10.18632/oncotarget.8559

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data Resource to drive high-quality survival outcome analytics. *Cell* 173 (2), 400–416. e411. doi:10.1016/j.cell.2018.02.052

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15 (12), 550. doi:10.1186/s13059-014-0550-8

Ma, W., Liang, J., Liu, J., Tian, D., and Chen, Z. (2021). Establishment and validation of an eight-gene metabolic-related prognostic signature model for lung adenocarcinoma. *Aging (Albany NY)* 13 (6), 8688–8705. doi:10.18632/aging.202681

Martinez, M., and Moon, E. K. (2019). CAR T cells for solid tumors: new strategies for finding, infiltrating, and surviving in the tumor microenvironment. *Front. Immunol.* 10, 128. doi:10.3389/fimmu.2019.00128

Mizuno, H., Kitada, K., Nakai, K., and Sarai, A. (2009). PrognoScan: a new database for meta-analysis of the prognostic value of genes. *BMC Med. Genomics* 2, 18. doi:10.1186/1755-8794-2-18

Modhukur, V., Iljasenko, T., Metsalu, T., Lokk, K., Laisk-Podar, T., and Vilo, J. (2018). MethSurv: a web tool to perform multivariable survival analysis using DNA methylation data. *Epigenomics* 10 (3), 277–288. doi:10.2217/epi-2017-0118

Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E., and Adjei, A. A. (2008). Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. *Mayo Clin. Proc.* 83 (5), 584–594. doi:10.4065/83.5.584

Ohta, M., Mimori, K., Fukuyoshi, Y., Kita, Y., Motoyama, K., Yamashita, K., et al. (2008). Clinical significance of the reduced expression of G protein gamma 7 (GNG7) in oesophageal cancer. *Br. J. Cancer* 98 (2), 410–417. doi:10.1038/sj.bjc.6604124

Petitprez, F., Meylan, M., de Reyniès, A., Sautès-Fridman, C., and Fridman, W. H. (2020). The tumor microenvironment in the response to immune checkpoint blockade therapies. *Front. Immunol.* 11, 784. doi:10.3389/fimmu.2020.00784

Qi, X., Qi, C., Qin, B., Kang, X., Hu, Y., and Han, W. (2020). Immune-Stromal score signature: novel prognostic tool of the tumor microenvironment in lung adenocarcinoma. *Front. Oncol.* 10, 541330. doi:10.3389/fonc.2020.541330

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., et al. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* 12, 77. doi:10.1186/1471-2105-12-77

Saito, M., Suzuki, H., Kono, K., Takenoshita, S., and Kohno, T. (2018). Treatment of lung adenocarcinoma by molecular-targeted therapy and immunotherapy. *Surg. Today* 48 (1), 1–8. doi:10.1007/s00595-017-1497-7

Schwindinger, W. F., Mirshahi, U. L., Baylor, K. A., Sheridan, K. M., Stauffer, A. M., Usefof, S., et al. (2012). Synergistic roles for G-protein γ3 and γ7 subtypes in seizure susceptibility as revealed in double knock-out mice. *J. Biol. Chem.* 287 (10), 7121–7133. doi:10.1074/jbc.M111.308395

Shibata, K., Mori, M., Tanaka, S., Kitano, S., and Akiyoshi, T. (1998). Identification and cloning of human G-protein gamma 7, down-regulated in pancreatic cancer. *Biochem. Biophys. Res. Commun.* 246 (1), 205–209. doi:10.1006/bbrc.1998.8581

Shibata, K., Tanaka, S., Shiraishi, T., Kitano, S., and Mori, M. (1999). G-protein gamma 7 is down-regulated in cancers and associated with p 27kip1-induced growth arrest. *Cancer Res.* 59 (5), 1096–1101.

Singh, S. S., Dahal, A., Shrestha, L., and Jois, S. D. (2020). Genotype driven therapy for non-small cell lung cancer: resistance, pan inhibitors and immunotherapy. *Curr. Med. Chem.* 27 (32), 5274–5316. doi:10.2174/0929867326666190222183219

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca. Cancer J. Clin.* 71 (3), 209–249. doi:10.3322/caac.21660

Tang, Z., Kang, B., Li, C., Chen, T., and Zhang, Z. (2019). GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Res.* 47 (W1), W556–w560. doi:10.1093/nar/gkz430

Taube, J. M., Galon, J., Sholl, L. M., Rodig, S. J., Cottrell, T. R., Giraldo, N. A., et al. (2018). Implications of the tumor immune microenvironment for staging and therapeutics. *Mod. Pathol.* 31 (2), 214–234. doi:10.1038/modpathol.2017.156

Tian, W., Zhou, J., Chen, M., Qiu, L., Li, Y., Zhang, W., et al. (2022a). Bioinformatics analysis of the role of aldolase A in tumor prognosis and immunity. *Sci. Rep.* 12 (1), 11632. doi:10.1038/s41598-022-15866-4

Tian, Y., Guan, L., Qian, Y., Wu, Y., and Gu, Z. (2022b). Effect of PPP1R14D gene high expression in lung adenocarcinoma knocked out on proliferation and apoptosis of DMS53 cell. *Clin. Transl. Oncol.* Online ahead of print. doi:10.1007/s12094-022-02842-7

Travis, W. D. (2011). Pathology of lung cancer. *Clin. Chest Med.* 32 (4), 669–692. doi:10.1016/j.ccm.2011.08.005

Wan, Q., Qu, J., Li, L., and Gao, F. (2021). Guanylate-binding protein 1 correlates with advanced tumor features, and serves as a prognostic biomarker for worse survival in lung adenocarcinoma patients. *J. Clin. Lab. Anal.* 35 (2), e23610. doi:10.1002/jcla.23610

Wu, S. G., and Shih, J. Y. (2018). Management of acquired resistance to EGFR TKI-targeted therapy in advanced non-small cell lung cancer. *Mol. Cancer* 17 (1), 38. doi:10.1186/s12943-018-0777-1

Wu, Z., Ouyang, C., and Peng, L. (2020). An immune scores-based nomogram for predicting overall survival in patients with clear cell renal cell carcinoma. *Med. Baltim.* 99 (34), e21693. doi:10.1097/md.0000000000021693

Xu, S., Zhang, H., Liu, T., Chen, Y., He, D., and Li, L. (2019). G Protein γ subunit 7 loss contributes to progression of clear cell renal cell carcinoma. *J. Cell. Physiol.* 234 (11), 20002–20012. doi:10.1002/jcp.28597

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4, 2612. doi:10.1038/ncomms3612

Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16 (5), 284–287. doi:10.1089/omi.2011.0118

Zhang, Y., Wang, Q., Zhu, T., and Chen, H. (2022). Identification of cigarette smoking-related novel biomarkers in lung adenocarcinoma. *Biomed. Res. Int.* 2022, 9170722. doi:10.1155/2022/9170722

Zhao, X., Zhang, X. C., Zang, K., and Yu, Z. H. (2021). MicroRNA miR-19b-3p mediated G protein γ subunit 7 (GNG7) loss contributes lung adenocarcinoma progression through activating Hedgehog signaling. *Bioengineered* 12 (1), 7849–7858. doi:10.1080/21655979.2021.1976896

Zheng, H., Tian, H., Yu, X., Ren, P., and Yang, Q. (2021). G protein gamma 7 suppresses progression of lung adenocarcinoma by inhibiting E2F transcription factor 1. *Int. J. Biol. Macromol.* 182, 858–865. doi:10.1016/j.ijbiomac.2021.04.082

Zhou, H., Yao, Y., Li, Y., Guo, N., Zhang, H., Wang, Z., et al. (2022). Identification of small nucleolar RNA SNORD60 as a potential biomarker and its clinical significance in lung adenocarcinoma. *Biomed. Res. Int.* 2022, 5501171. doi:10.1155/2022/5501171

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 01 November 2022
DOI 10.3389/fgene.2022.1032010

EDITED BY
Saurav Mallik,
Harvard University, United States

REVIEWED BY
Wafa Hamdi,
Tunis El Manar University, Tunisia
Shima Sadri,
Marquette University, United States
Partho Ghose,
Bangladesh University of Business and Technology, Bangladesh

*CORRESPONDENCE
Xieping Dong,
jxsrmyydongxieping@163.com

†These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 30 August 2022
ACCEPTED 14 October 2022
PUBLISHED 01 November 2022

CITATION
Wen J, Wan L and Dong X (2022), Novel peripheral blood diagnostic biomarkers screened by machine learning algorithms in ankylosing spondylitis.
Front. Genet. 13:1032010.
doi: 10.3389/fgene.2022.1032010

Novel peripheral blood diagnostic biomarkers screened by machine learning algorithms in ankylosing spondylitis

Jian Wen[1,2†], Lijia Wan[3†] and Xieping Dong[1,2*]

[1]Medical College of Nanchang University, Nanchang, Jiangxi, China, [2]JXHC Key Laboratory of Digital Orthopedics, Department of Orthopedics, Jiangxi Provincial People's Hospital, The First Affiliated Hospital of Nanchang Medical College, Nanchang, Jiangxi, China, [3]Department of Child Healthcare, Hunan Provincial Maternal and Child Health Hospital, Changsha, Hunan, China

Background: Ankylosing spondylitis (AS) is a chronic inflammatory disorder of unknown etiology that is hard to diagnose early. Therefore, it is imperative to explore novel biomarkers that may contribute to the easy and early diagnosis of AS.

Methods: Common differentially expressed genes between normal people and AS patients in GSE73754 and GSE25101 were screened by machine learning algorithms. A diagnostic model was established by the hub genes that were screened. Then, the model was validated in several data sets.

Results: *IL2RB* and *ZDHHC18* were screened using machine learning algorithms and established as a diagnostic model. Nomograms suggested that the higher the expression of *ZDHHC18*, the higher was the risk of AS, while the reverse was true for *IL2RB in vivo*. C-indexes of the model were no less than 0.84 in the validation sets. Calibration analyses suggested high prediction accuracy of the model in training and validation cohorts. The area under the curve (AUC) values of the model in GSE73754, GSE25101, GSE18781, and GSE11886 were 0.86, 0.84, 0.85, and 0.89, respectively. The decision curve analyses suggested a high net benefit offered by the model. Functional analyses of the differentially expressed genes indicated that they were mainly clustered in immune response–related processes. Immune microenvironment analyses revealed that the neutrophils were expanded and activated in AS while some T cells were decreased.

Conclusion: *IL2RB* and *ZDHHC18* are potential blood biomarkers of AS, which might be used for the early diagnosis of AS and serve as a supplement to the existing diagnostic methods. Our study deepens the insight into the pathogenesis of AS.

Abbreviations: AS, ankylosing spondylitis; AUC(s), area under curve(s); BP, biological process; CC, cellular component; CI, confidence interval; DCA, decision curve analysis; DEGs, differentially expressed genes; GEO, Gene Expression Omnibus database; GO, Gene ontology; GSEA, Gene Set Enrichment Analysis; HPA, Human Protein Atlas; KEGG, Kyoto Encyclopedia of Genes and Genomes; KM, Kaplan–Meier; LASSO, least absolute shrinkage and selection operator; MF, molecular function; OS, overall survival; PPI, protein–protein interaction; RF, random forest; ROC, receiver operating characteristic; Tregs, regulatory T cells.

KEYWORDS

machine learning, ankylosing spondylitis, diagnostic model, immune microenvironment, informatics

Introduction

Ankylosing spondylitis (AS), also known as radiographic axial spondyloarthritis, is one of the two types of axial spondyloarthritides (Sieper et al., 2015; Taurog et al., 2016; Sieper and Poddubnyy, 2017; Navarro-Compán et al., 2021). It is a chronic inflammatory disorder mainly affecting the axial joints and entheses and is usually characterized by typical features such as inflammatory back pain, limitation of the motion of the lumbar spine, restricted chest expansion, and advanced sacroiliitis on plain radiographs. Some patients with AS also experience peripheral spondyloarthritis symptoms such as dactylitis and Achilles tendinitis and extra-articular manifestations such as uveitis, psoriasis, inflammatory bowel disease, and many others, either simultaneously or at some point during the course of the disease. The diagnosis of AS is based on the Modified New York criteria: advanced sacroiliitis on plain radiographs with any one of the three typical aforementioned features (van der Linden et al., 1984). Patients usually do not meet the criterion of advanced sacroiliitis on plain radiographs; however, those with sacroiliitis on MRI or HLA-B27 positivity plus the clinical criteria are classified into non-radiographic axial spondyloarthritis (Rudwaleit et al., 2009; Rudwaleit et al., 2011).

The prevalence of AS, which reportedly varies with geography, ranges from 0.02–0.35%, while that of axial spondyloarthritis is estimated to be 0.20–1.61%, which is much higher than the prevalence of AS, indicating a high ratio of non-radiographic axial spondyloarthritis patients (Dean et al., 2014; Stolwijk et al., 2016; Ward et al., 2019). Especially, with the development of diagnostic tools and further understanding of axial spondyloarthritis, patients without advanced sacroiliitis on plain radiographs raise more attention, and more non-radiographic axial spondyloarthritides are detected together with updates in its definition (Taurog et al., 2016; Ritchlin and Adamopoulos, 2021). However, even with modern diagnostic methods, the diagnostic sensitivity and specificity for axial spondyloarthritis are not higher than approximately 80% (Sieper and Poddubnyy, 2017). This means that a significant number of patients are still excluded from the current diagnostic criteria, and there is still a lot of room for improvement in our diagnostic methods. More importantly, it has been reported that approximately 10–20% of patients with non-radiographic axial spondyloarthritis will progress to AS within 1 year after the initial diagnosis while 20.3% of them will do so in 2–6 years (Sieper and Poddubnyy, 2017). Therefore, it is necessary to identify pre-AS patients, for identifying them could save more time for clinical interventions.

At present, our measures to identify axial spondyloarthritis are still limited beyond clinical features. Imaging (radiography, CT, and MRI), HLA-B27, and C-reactive protein (CRP) features are the main indices for the clinical diagnosis of axial spondyloarthritis (Zochling et al., 2005; Sieper and Poddubnyy, 2017; Ritchlin and Adamopoulos, 2021). More methods with high sensitivity and specificity are eagerly expected. Although with the rapid development of genomics technology, many serum biomarkers for the diagnosis of AS such as miR-214 (Kook et al., 2019), deoxyribonuclease 1-like 3 (Sun et al., 2020), anti-SIRT1 autoantibody (Hu et al., 2018), sclerostin (Perrotta et al., 2018), endoplasmic reticulum aminopeptidase 1 (Danve and O'Dell, 2015), and others have been identified, there is still a paucity of reliable indices for clinical practice besides HLA-B27 and CRP (Sieper and Poddubnyy, 2017; Danve and O'Dell, 2015). Therefore, the exploration of gene biomarkers of AS in peripheral blood is not only of real need and great practical value but could also deepen our knowledge of the pathophysiology of AS and even help us understand its etiology.

Thereby, in this study, we aimed to screen potential gene biomarkers in the peripheral blood by machine learning algorithms and build a diagnostic model and also preliminarily explore the immune microenvironment of AS to find some differences in immune cell proportions and potential explanations for our hub genes. To date, this work has not been done and reported; thus, it is imperative to bridge the knowledge gap in this area.

Materials and methods

Data collection

We searched the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) for data sets containing whole-blood RNA expression data of normal people and AS patients with at least 15 samples in each group. Only GSE73754, GSE25101, and GSE18781 were qualified, and their expression and phenotype data were downloaded for subsequent studies. GSE73754 and GSE18781 contained whole-blood RNA expression data of 20 normal and 52 AS patients and 25 normal and 18 AS patients, respectively, together with their corresponding basic information such as sex and age. The expression data of GSE73754 were detected by the Illumina HumanHT-12 V4.0 expression BeadChip, University of Toronto, Canada, submitted on 06 Oct 2015. The expression data of GSE18781 were detected by the Affymetrix Human Genome U133 Plus 2.0 Array, Oregon Health & Science University, United States, submitted on 28 Oct 2009. GSE25101 contained whole-blood RNA expression data of 16 normal and 16 AS patients, which were detected by the

Illumina HumanHT-12 V3.0 expression BeadChip, University of Queensland Diamantina Institute, Australia, submitted on 03 Nov 2010. However, the basic information of the subjects from GSE25101 was unavailable; so, it is only used as one of the validation sets. GSE11886 referred to the RNA expression data of *in vitro* cultured macrophages, which were obtained from the peripheral blood of nine normal people and eight AS patients. They were detected by the Affymetrix Human Genome U133 Plus 2.0 Array, Cincinnati Children's Hospital Medical Center, United States, submitted on 25 Jun 2008. Although the RNA expression data of each set were normalized data, while in the quality control process, we found that samples of GSE18781 came from two batches; so, we used the "removebatcheffect" function of the "limma" package to recalculate the expression data (Ritchie et al., 2015).

Identify common differentially expressed genes

Differentially expressed genes (DEGs) in GSE73754 and GSE25101 between normal people and AS patients were identified by the "limma" package (Ritchie et al., 2015) (cutoff value: the absolute value of $\log_2$foldchange >0.3 and p-value < 0.05). Then, common DEGs in GSE73754 and GSE25101 were selected as candidates for subsequent screening.

Screening genes for diagnostic model by machine learning algorithms

GSE73754 served as the training set. Common DEGs were first screened by univariate logistic regression in the training set. Genes with a p-value < 0.05 were retained. Then, three machine learning algorithms: the least absolute shrinkage and selection operator (LASSO) logistic regression (Simon et al., 2011), a support vector machine recursive feature elimination (SVM-RFE) (Sanz et al., 2018), and random forest (RF) (Strobl et al. 2007) were adopted to screen hub genes. The common hub genes were selected as the final genes for the diagnostic model.

Establishment of diagnostic model and its evaluation in training set and related validation set

A diagnostic model was established by the common hub genes and visualized by nomograms. Then, the prediction accuracy and discriminatory capacity were first assessed in GSE73754 and GSE25101 by the C-index, calibration analysis, receiver operating characteristic (ROC) curves, and decision curve analysis (DCA).

Validation of model in validation sets

GSE18781 was set as an *in vivo* external validation set, while GSE11886 was set as an *in vitro* external validation set. The prediction accuracy and discriminatory capacity of the model were also assessed in the two aforementioned sets by the C-index, calibration analysis, ROC analysis, and DCA.

Functional analysis of differentially expressed genes between normal and ankylosing spondylitis groups

GO and KEGG clustering and gene set enrichment analyses (GSEA) were used to explore the potential functions of the DEGs, which might indicate the causes of the difference between normal people and AS patients. With the same consideration, the protein–protein interaction (PPI) network analysis was also adopted to investigate the interaction between the proteins encoded by the DEGs (interaction score ≥0.4).

Immune microenvironment analysis

The "CIBERSORT" package was employed to investigate the immune microenvironment (IME) of the samples. Meanwhile, the correlations between the different types of immune cells and the hub genes were also explored.

Statistical analyses

In this study, the R software v3.63 was used to process data and generate charts. PPI network analyses were explored on the STRING website (https://cn.string-db.org/) (interaction score ≥0.4) and visualized by the Cytoscape software v3.7.1. Flexible statistical methods were adopted for the statistical analyses.

Results

Clinical characteristics of enrolled ankylosing spondylitis patients

The basic information of the samples from GSE73754 and GSE18781 is shown in Table 1. The clinical characteristics such as age and sex of the two sets were similar (p-value < 0.05).

Identification of hub genes

In total, 64 downregulated and 132 upregulated DEGs were identified by "limma" in GSE73754 (Figure 1A). Also,

TABLE 1 Clinical characteristics in training and validation sets.

Characteristics	Level	GSE18781	GSE73754	*p*-value	Test
Sample size (n)		43	72		
Sex	Female	25 (58.1)	35 (48.6)	0.342	Fisher test
	Male	18 (41.9)	37 (51.4)		
Age, median (interquartile range)		45.0 [32.5, 58.5]	41.5 [28.8, 51.2]	0.324	Kruskal test
Group	Normal	25 (58.1)	20 (27.8)		
	AS	18 (41.9)	52 (72.2)		

278 downregulated and 345 upregulated DEGs were identified in GSE25101 (Figure 1B). Then, the common upregulated and downregulated genes were selected: three common downregulated genes, namely, *IL2RB*, *GZMM*, and *CXXC5* (Figure 1C), and four common upregulated genes, namely, *S100A12*, *ANXA3*, *PROS1*, and *ZDHHC18* (Figure 1D).

Taking GSE73754 as the training set, the *p*-values of the seven genes in the univariate logistic regression were all lower than 0.05, meaning that all seven genes were qualified for the next screening. Then, they were screened by three different machine learning algorithms. *IL2RB*, *GZMM*, *S100A12*, and *ZDHHC18* were screened as hub genes by LASSO (λ = lambda.min) (Figures 1E,F). *IL2RB* and *ZDHHC18* were screened as hub genes by SVM-RFE (Figure 1G). *ZDHHC18*, *CXXC5*, *PROS1*, and *IL2RB* were screened as hub genes by RF with MeanDecreaseAccuracy >3 and MeanDecreaseGini >2 (mtry = 3, ntree = 200) (Figures 1H,I). Obviously, *IL2RB* and *ZDHHC18* were the common hub genes screened by the three algorithms, and they were selected as the final hub genes for a diagnostic model in AS.

Evaluation of diagnostic model in training set (GSE73754) and GSE25101

A diagnostic model was established by *IL2RB* and *ZDHHC18* and then visualized by a nomogram in GSE73754 (Figure 2A) and GSE25101 (Figure 2B), respectively. The nomograms suggested that the higher the expression level of *ZDHHC18* was, the higher was the risk of AS, while the reverse was true for *IL2RB*. The C-index of the diagnostic model in GSE73754 was 0.86 (95% CI: 0.76–0.96) and 0.84 (95% CI: 0.71–0.97) in GSE25101. The calibration analyses showed that the predicted probability was in high agreement with the observed probability, suggesting a high accuracy of the model both in the training set and an external validation set (Figures 2C,D).

The ROC analysis in GSE73754 showed that the areas under the curves (AUCs) for the nomogram, *IL2RB*, and *ZDHHC18* were 0.86, 0.83, and 0.83, respectively (Figure 2E). The optimal truncation value of Y was 0.713, and the corresponding specificity and sensitivity were 0.85 and 0.827, respectively (formula: $y = 2.9111*EXP_{ZDHHC18} - 2.3256*EXP_{IL2RB} - 2.2376$, where $EXP_{ZDHHC18}$ refers to the expression value of *ZDHHC18* and EXP_{IL2RB} refers to the expression value of *IL2RB*). In this model, the value of Y ≥ 0.713, predicted to be AS, was otherwise normal. The actual prediction accuracy of the model in GSE73754 was 0.82. While in GSE25101, the AUCs for the nomogram, *IL2RB*, and *ZDHHC18* were 0.84, 0.79, and 0.76, respectively (formula: $y = 2.320{,}052*EXP_{ZDHHC18} - 1.728{,}388*EXP_{IL2RB} - 6.902{,}309$) (Figure 2F). There were three optimal truncation values for Y: 0.589 with a corresponding specificity of 0.875 and sensitivity of 0.688, 0.521 with a corresponding specificity of 0.75 and sensitivity of 0.812, and 0.452 with a corresponding specificity of 0.688 and sensitivity of 0.875. The actual prediction accuracy of the model in GSE25101 was 0.72. The DCA for the nomogram and models involved only one of these genes, which indicated that the net benefit of the nomogram was higher than that of the other models (Figures 2G,H).

Validation of model in independent cohort and *in vitro*

The model was validated in an independent cohort, GSE18781, and *in vitro* cohort, GSE11886. The nomogram for GSE18781 supported the conclusion reached in the training set that AS patients had a higher expression of *ZDHHC18* and lower expression of *IL2RB* (Figure 3A). The function of *IL2RB* in GSE11886 was in accordance with that in the other sets; however, the function of *ZDHHC18 in vitro* was opposite to that *in vivo*, and this might have been due to the lack of the *in vivo* microenvironment (Figure 3B). According to the coverage of points in the nomogram, *IL2RB* showed higher weight in the validation sets and the alteration between the nomograms also indicated that it is a more robust indicator than *ZDHHC18*. The C-index of the diagnostic model in GSE18781 was 0.85 (95% CI: 0.73–0.96) and 0.89 (95% CI: 0.73–1.05) in GSE11886. The calibration analyses revealed that the prediction accuracy of the model was lower than that in GSE73754 and GSE25101; however, it still had acceptable accuracy (Figures 3C,D).

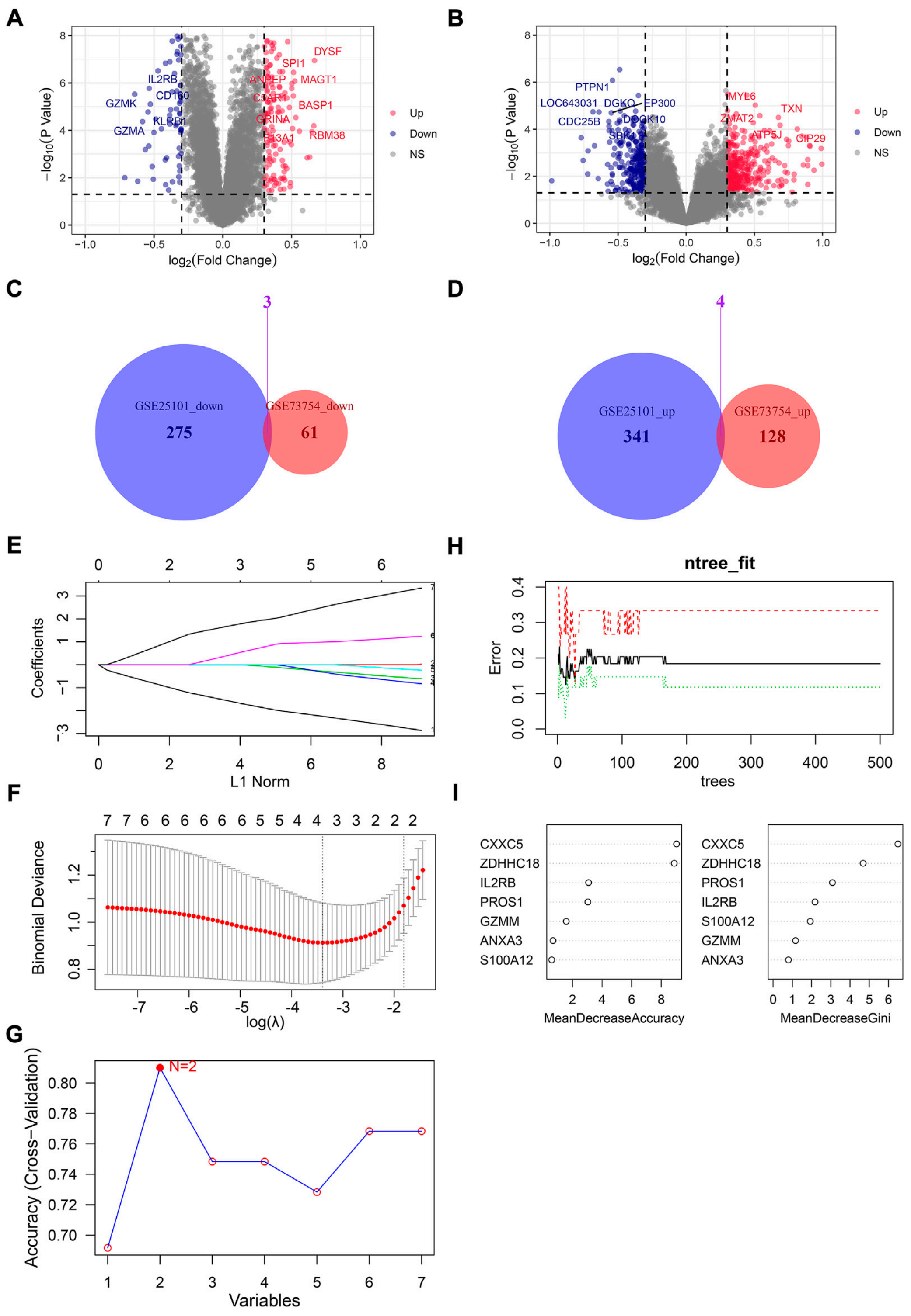


FIGURE 1
Screening for hub genes from DEGs between normal people and AS patients. The volcano plot for DEGs in GSE73754 **(A)** and GSE25101 **(B)**: *x*-axis represents $\log_2$ (fold change) of gene expressions in AS patients compared with normal controls, while the *y*-axis represents $-\log_{10}$ (*p*-value) of gene expression between AS patients and normal controls. **(C)** Venn plot for downregulated DEGs in GSE73754 and GSE25101. **(D)** Venn plot for upregulated DEGs in GSE73754 and GSE25101. **(E)** LASSO coefficient profiles for the seven common DEGs in the ten-fold cross-validations. **(F)** Partial likelihood deviance with changing of log(λ) plotted by LASSO regression in ten-fold cross-validations. **(G)** Filtering characteristic genes using the SVM-RFE algorithm: accuracy for models with different numbers of variables: the *x*-axis represents the number of variables involved in the models and the *y*-axis represents the corresponding accuracy of cross-validation of the models. **(H)** Relationship between the number of decision trees and the error rate of the model in RF. **(I)** Selecting hub genes by variable importance measures for RF.

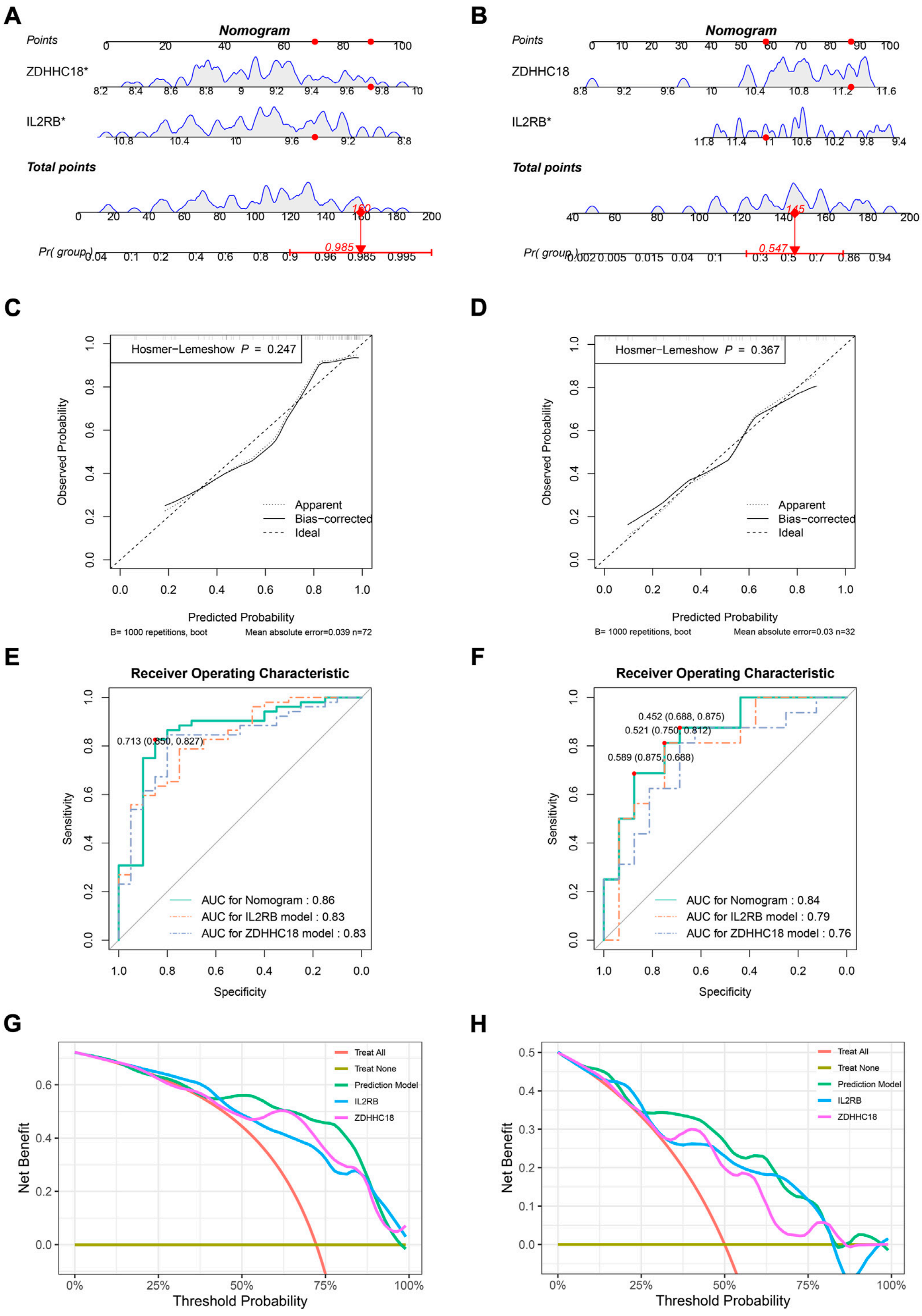


FIGURE 2
Evaluating the diagnostic model in the training set and a related validation set. Nomograms for the diagnostic model in GSE73754 **(A)** and GSE25101 **(B)**. Calibration plots for the diagnostic model in GSE73754 **(C)** and GSE25101 **(D)**: *x*-axis represents the predicted probability of AS by the model, while the *y*-axis represents the observed probability of AS, the diagonal (dashed line) represents the ideal status that the predicted probability equaled the observed probability, and the solid and dotted lines represent the apparent and bias-corrected statuses of the predicted and observed probabilities, respectively. ROC plots for the diagnostic model in GSE73754 **(E)** and GSE25101 **(F)**: the *x*-axis represents 1-specificity of the model, while the *y*-axis represents the sensitivity of the model. DCA in GSE73754 **(G)** and GSE25101 **(H)**: the *x*-axis represents the threshold probability for the treatment or intervention, while the *y*-axis represents the net benefit.

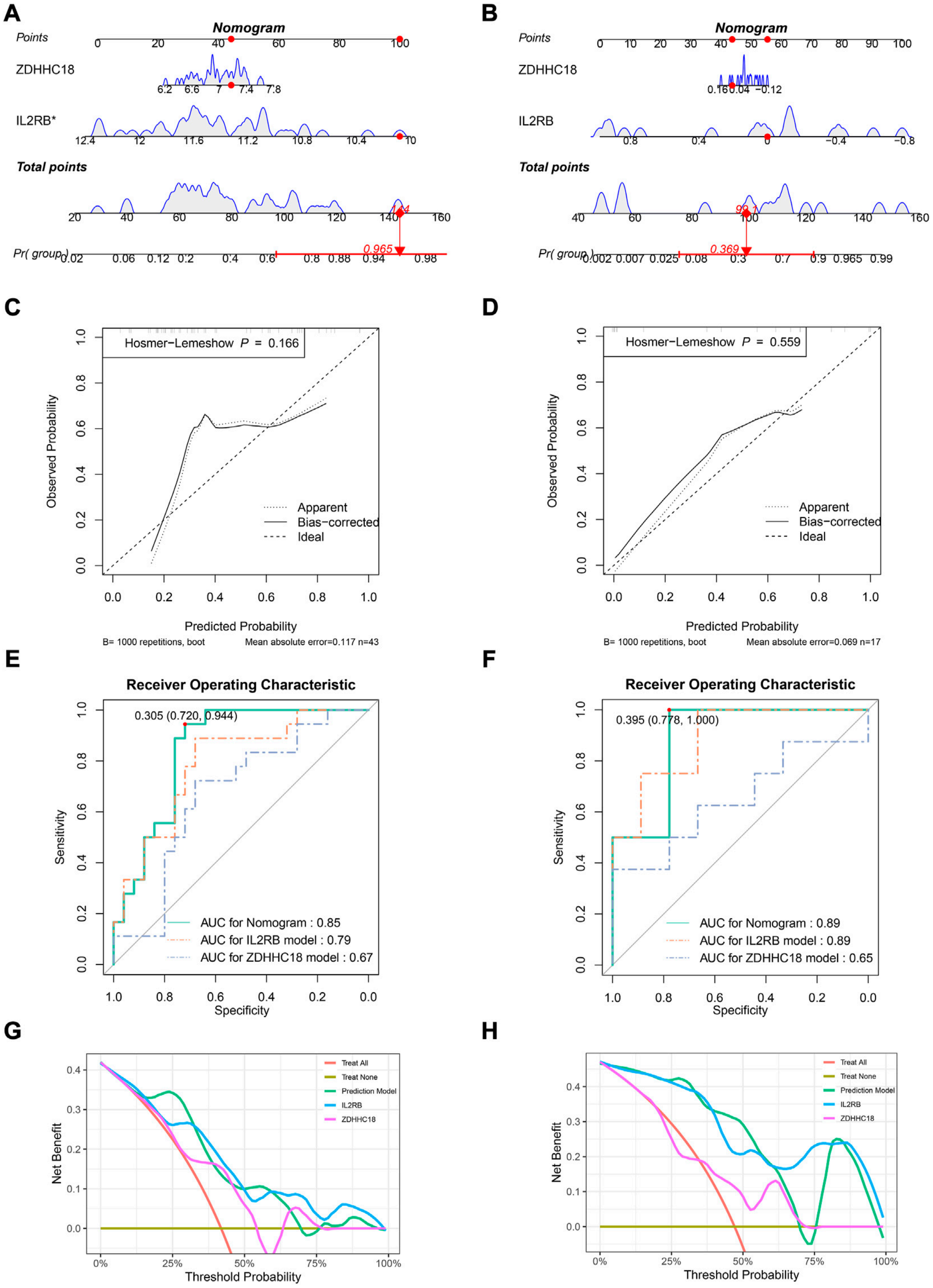
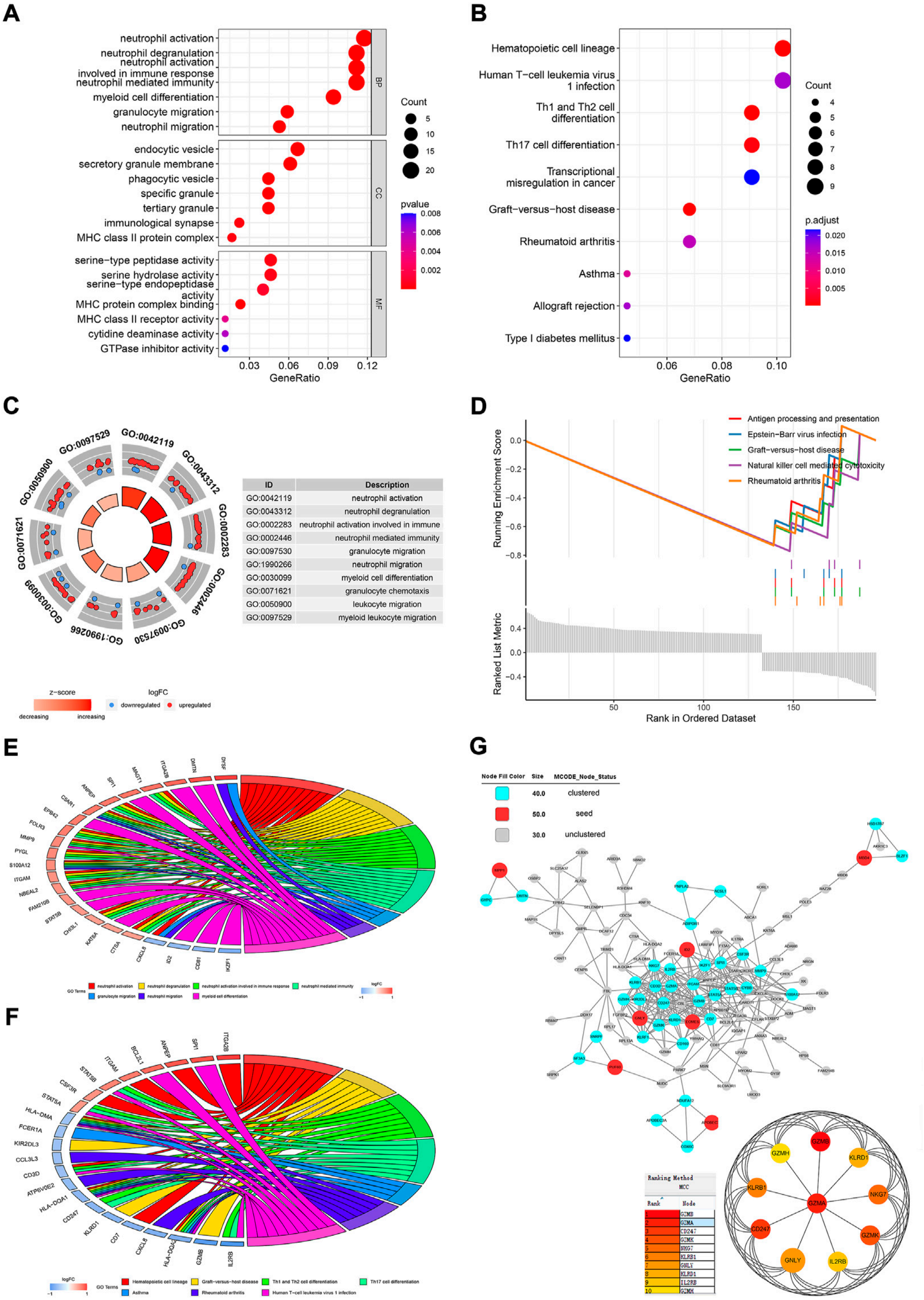


FIGURE 3
Validating the diagnostic model in validation sets. Nomograms for the diagnostic model in GSE18781 **(A)** and GSE11886 **(B)**. Calibration plots for the diagnostic model in GSE18781 **(C)** and GSE11886 **(D)**: the *x*-axis represents the predicted probability of AS by the model, while the *y*-axis represents the observed probability of AS. The diagonal (dashed line) represents the ideal status that the predicted probability equaled the observed probability, and the solid and dotted lines represent the apparent and bias-corrected statuses of the predicted and observed probabilities, respectively. ROC plots for the diagnostic model in GSE18781 **(E)** and GSE11886 **(F)**: the *x*-axis represents 1-specificity of the model, while the *y*-axis represents the sensitivity of the model. DCA in GSE18781 **(G)** and GSE11886 **(H)**: the *x*-axis represents the threshold probability of the treatment or intervention, while the *y*-axis represents the net benefit.



ID	Description
GO:0042119	neutrophil activation
GO:0043312	neutrophil degranulation
GO:0002283	neutrophil activation involved in immune
GO:0002446	neutrophil mediated immunity
GO:0097530	granulocyte migration
GO:1990266	neutrophil migration
GO:0030099	myeloid cell differentiation
GO:0071621	granulocyte chemotaxis
GO:0050900	leukocyte migration
GO:0097529	myeloid leukocyte migration

Ranking Method
MCC

Rank	Node
1	GZMB
2	GZMA
3	CD247
4	GZMK
5	NKG7
6	KLRB1
7	GNLY
8	KLRD1
9	IL2RB
10	GZMH

FIGURE 4
Functional analysis of the DEGs between normal people and AS patients. Dot plots for GO **(A)** and KEGG **(B)** analyses of DEGs. **(C)** Circle plot for BP clustering of the DEGs. **(D)** GSEA analysis for the DEGs. **(E)** Chord plot for the top seven clustered GO terms. **(F)** Chord plot for the top seven clustered KEGG pathways. **(G)** PPI network analysis for DEGs.

The ROC analysis in GSE18781 revealed that the areas under the curves (AUCs) for the nomogram, *IL2RB*, and *ZDHHC18* were 0.85, 0.79, and 0.67, respectively (Figure 3E). The optimal truncation value of Y was 0.305, and the corresponding specificity and sensitivity were 0.72 and 0.994, respectively (formula: y = 1.29499*EXP$_{ZDHHC18}$ − 2.582,298*EXP$_{IL2RB}$ + 20.055204). The actual prediction accuracy of the model in GSE18781 was 0.72. While in GSE11886, the AUCs for nomogram, *IL2RB*, and *ZDHHC18* were 0.89, 0.89, and 0.65, respectively (formula: y = −6.49159*EXP$_{ZDHHC18}$ − 6.13506*EXP$_{IL2RB}$ − 0.01334) (Figure 3F). The optimal truncation value of Y was 0.395, and the corresponding specificity and sensitivity were 0.778 and 1, respectively. The actual prediction accuracy of the model in GSE11886 was 0.76. The DCA showed that patients could get a high net benefit from the nomogram (Figures 3G,H). Besides, a high net benefit could also be obtained from the model established by IL2RB only in this set.

Results of functional analysis of differentially expressed genes between normal and ankylosing spondylitis groups

There was a total of 196 DEGs between normal people and AS patients in GSE73754. Biological process (BP) clustering of the DEGs showed that they were mainly clustered in neutrophil activation, degranulation, immune response, and migration (Figure 4A). Myeloid cell differentiation, leukocyte migration, and granulocyte migration were also clustered BPs. Gene clustering of cellular components (CC) was mostly in the area of membranes, such as endocytic vesicles, secretory granule membranes, membrane microdomains, and cytoplasmic vesicle lumens (Figure 4A). Molecular functions (MFs) of the DEGs were mostly clustered in serine-type peptidase activity, serine hydrolase activity, serine-type endopeptidase activity, and MHC protein complex binding (Figure 4A). In the KEGG clustering of the DEGs, the hematopoietic cell lineage, human T-cell leukemia virus 1 infection, Th1 and Th2 cell differentiation, and Th17 cell differentiation were the top clustered pathways (Figure 4B). The circle plot for BP clustering showed that neutrophil activation, degranulation, immune response, and migration were upregulated in AS (Figure 4C). By GSEA, antigen processing and presentation, natural killer cell–mediated cytotoxicity, graft-*versus*-host disease, Epstein–Barr virus infection, and rheumatoid arthritis were the top enriched gene sets, which were all downregulated in AS patients (Figure 4D). The top three upregulated pathways enriched with core enrichment genes were neutrophil extracellular trap formation, complement and coagulation cascades, and the rap1 signaling pathway. The GO chord plot showed that *DYSF*, *DMTN*, *ITGA2B*, *MAGT1*, *SPI1*, *CXCL8*, *ID2*, *CD81*, *IKZF1*, and many others were involved in the top seven GO terms (Figure 4E). The KEGG chord plot showed that *ITGA2B*, *SPI1*, *ANPEP*, *BCL2L1*, *STAT5B*, *IL2RB*, *GZMB*, *HLA-DQA2*, *CXCL8*, and many more were involved in the top seven KEGG terms (Figure 4F).

The PPI network of the proteins encoded by DEGs showed that MMP1, ID2, MBD4, GNLY, EOMES, PUF60, and APOBEC3G were seed proteins in the network by the MCODE application in Cytoscape (Figure 4G). The cyan nodes were also pivotal nodes in the net, such as IL2RB, GZMA, SPI1, and many others. Then, GZMA, IL2RB, CD247, KLRB1, GZMH, GZMB, GZMK, KLRD1, NKG7, and GNLY were the top 10 hub proteins screened by cytoHubba.

Results of immune microenvironment analyses

IME analyses were performed in GSE73754, GSE25101, and GSE18781 by CIBERSORT. The proportions of the 22 immune cells for samples are shown in Figures 5A–C. In all three sets, the neutrophils and monocytes accounted for the top two highest proportions and together made up the majority of the immune cells, while the other immune cells such as granulocytes, B cells, dendritic cells, and macrophages each made up only a small proportion of the total immune cell population. The relative quantities of different immune cells in normal people and AS patients are shown in Figures 5D–F. In GSE73754, when compared with the normal subjects, there were more neutrophils and naive CD4 T cells detected in the blood of AS patients, while there were fewer resting NK, CD8$^+$ T, and gamma-delta T cells (Figure 5D). In GSE25101, monocytes were found to be more in the blood of AS patients, while regulatory T cells (Tregs) were fewer. In this set, the relative number of neutrophils was more in the AS group; however, the difference was not statistically significant (Figure 5E). The result in GSE18781 was similar to that in GSE73754; the relative number of neutrophils was increased, while that of CD8$^+$ and gamma-delta T cells was decreased in patients with AS (Figure 5F).

The correlation between our hub genes (*IL2RB* and *ZDHHC18*) and immune cells was also explored. In GSE73754, the expression of *IL2RB* was positively correlated with the relative numbers of resting NK, CD8$^+$ T, and gamma-delta cells (Figures 6A–C) and negatively correlated with the relative numbers of neutrophils, naive CD4 T cells, and monocytes (Figures 6D,E). Meanwhile, the expression of *ZDHHC18* was positively correlated with the relative number of neutrophils (Figure 6F) but negatively correlated with the relative numbers of CD8$^+$ T cells and resting NK cells (Figures 6G,H). In GSE25101, the expression of *IL2RB* was positively correlated with the relative numbers of resting NK and activated

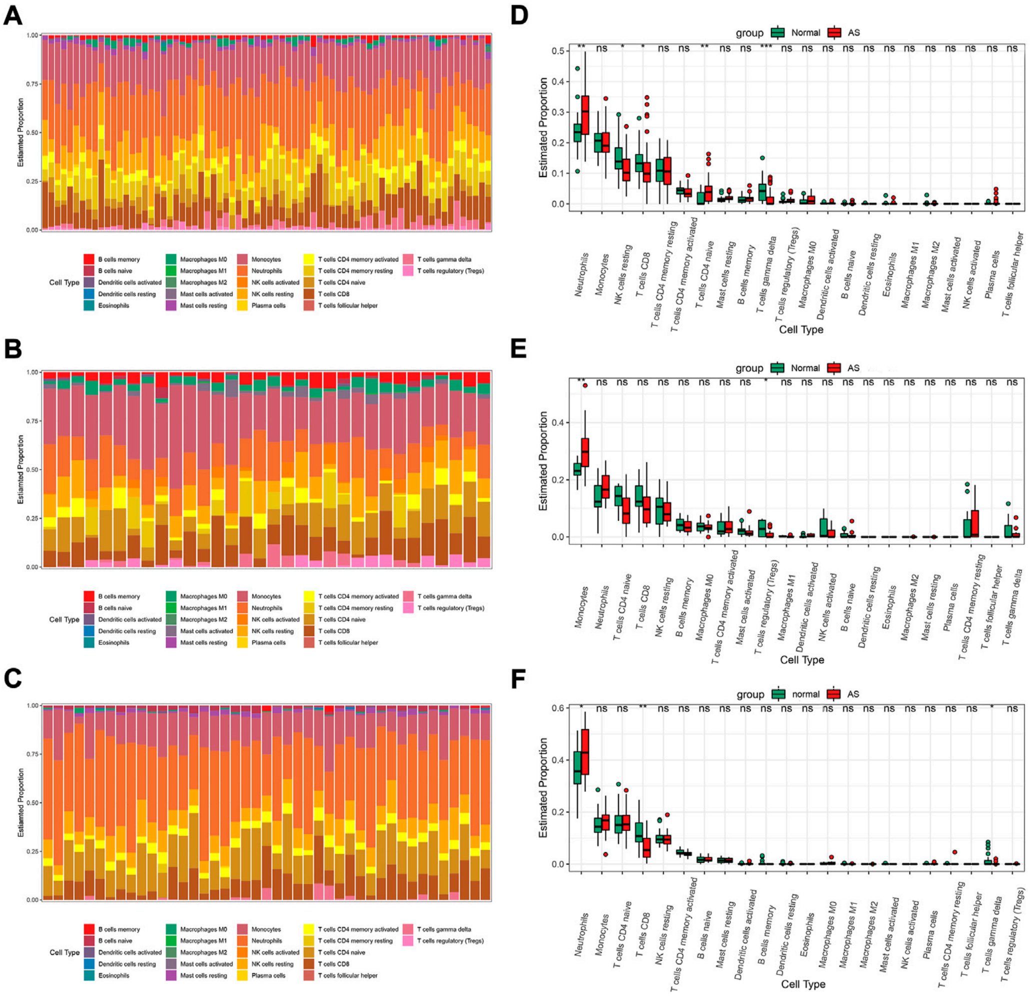


FIGURE 5
The IME analysis of the sets by CIBERSORT. The proportion of the 22 immune cells for samples in GSE73754 **(A)**, GSE25101 **(B)**, and GSE18781 **(C)**. Boxplots for the 22 immune cells between normal people and AS patients in GSE73754 **(D)**, GSE25101 **(E)**, and GSE18781 **(F)** (p significance level: no significance (ns), $p \geq 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.).

$CD4^+$ memory T cells (Figures 6I,J) and negatively correlated with the relative number of monocytes (Figure 6K). Besides, the expression of *ZDHHC18* was positively correlated with the relative number of neutrophils (Figure 6L) but negatively correlated with the relative number of activated NK cells (Figure 6M). There was no significant correlation between Tregs and the hub genes. Lastly, in GSE18781, the expression of *IL2RB* was positively correlated with the relative numbers of resting NK and $CD8^+$ T cells (Figures 6N,O) and negatively correlated with the relative number of neutrophils (Figure 6P). Moreover, the expression of *ZDHHC18* was positively correlated with the relative number of neutrophils (Figure 6Q) but negatively correlated with the relative quantities of $CD8^+$, gamma-delta, and activated $CD4^+$ memory T cells (Figures 6R–T).

Discussion

It is known that AS is an inflammatory disease mainly involving the axial skeleton's joints and entheses. The essential change in AS is the dysregulation of inflammation by innate and adaptive immune responses (Mauro et al., 2021). Although AS is primarily associated with the axial skeleton, recent research indicates that it may be initiated in the gut (Yang et al., 2016a). Besides, the peripheral and extra-articular manifestations of AS also suggest that it is a systemic disorder. Therefore, DEGs in the peripheral blood of AS patients can also reflect some features of AS. As for RNAs extracted from the peripheral blood, they are mostly from the nucleated cells in the blood, similar to leukocytes and immature red blood cells; so, it is rational to explore the immune

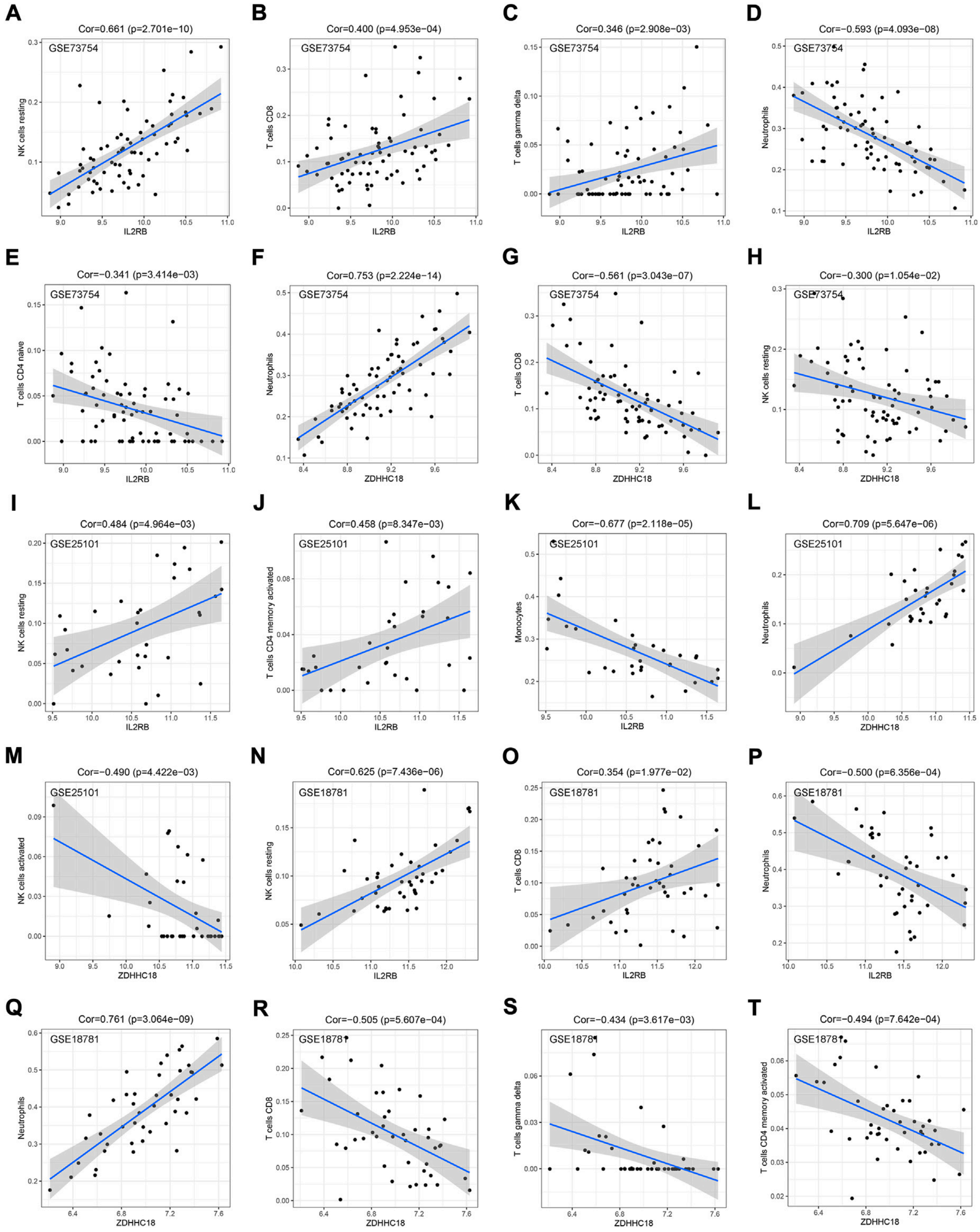


FIGURE 6

Correlation between hub genes and IME cells. The correlation between the expression of IL2RB and the estimated proportion of resting NK cells **(A)**, CD8^{+} T cells **(B)**, gamma-delta T cells **(C)**, neutrophils **(D)**, and native CD4 T cells **(E)** by CIBERSORT in GSE73754. The correlation between the expression of ZDHHC18 and the estimated proportions of neutrophils **(F)**, CD8^{+} T cells **(G)**, and resting NK cells **(H)** by CIBERSORT in GSE73754. The correlation between the expression of IL2RB and the estimated proportions of resting NK cells **(I)**, activated CD4^{+} memory T cells **(J)**, and monocytes **(K)** by CIBERSORT in GSE25101. The correlation between the expression of ZDHHC18 and the estimated proportions of neutrophils **(L)** and activated NK cells **(M)** by CIBERSORT in GSE25101. The correlation between the expression of IL2RB and the estimated proportions of resting NK cells **(N)**, CD8^{+} T cells **(O)**, and neutrophils **(P)** by CIBERSORT in GSE18781. The correlation between the expression of ZDHHC18 and the estimated proportions of neutrophils **(Q)**, CD8^{+} T cells **(R)**, gamma-delta T cells **(S)**, and activated CD4^{+} memory T cells **(T)** by CIBERSORT in GSE18781.

microenvironment of the blood of AS patients. More importantly, compared with the focal tissue, the peripheral blood is easier to obtain and a more commonly used clinical detection material, which is also conducive to the transition from experimental results to applications.

To date, HLA-B27 is still considered the most important factor in the pathogenesis of AS (Colbert et al., 2010; Bowness, 2015; Pedersen and Maksymowych, 2019; Sharip and Kunz, 2020; Voruganti and Bowness, 2020). First, many shreds of evidence supported the hypothesis that the alternation of the amino acid sequence in the antigenic peptide-binding groove of HLA-B27 might induce changes in the binding specificity of peptides and result in CD8$^+$ T cell–mediated immune cross-reactivity in the AS focus (Mear et al., 1999; Guiliano et al., 2017). Second, endoplasmic reticulum stress was induced by the accumulation of misfolded HLA-B27, which led to an unfolded protein response (UPR) and autophagy (Yu et al., 2017). Third, the HLA-B27 homodimer hypothesis suggests that the HLA-B27 homodimer could activate CD8$^+$ T cells and NK cells by the specific receptors on their surfaces, activating the IL-23/IL-17 axis (Bowness et al., 2011). Certainly, there were also many other hypotheses, such as the non-MHC hypothesis. However, the point of intersection is that all the hypotheses are focused on the antigen-presenting process, and its failure or dysfunction would mostly result in the activation of the TNF signaling pathway and the IL23/IL17 axis and eventually lead to the AS phenotype. However, the sensitivity and specificity of HLA-B27 alone were relatively low.

Here, to enhance the reliability and stability of the results, only common genes screened by the three machine learning algorithms were selected as hub genes for a diagnostic model. The three methods used in our study are the most popular and widely used ones in bioinformatics analyses. Currently, deep learning methods are also popular in bioinformatics analyses, and some of them can even generate different methods based on machine learning techniques such as BioSeq-BLM and ilearn. However, they are limited by the quantity and quality of the training data and are more suitable for large data processing (Choi et al., 2020). The data used in our study are small; so, the three machine learning methods could be more suitable. Meanwhile, deep learning methods are more complex, time-consuming, have high requirements for computer hardware, and have results that are more difficult to interpret (Choi et al., 2020). Besides, we validated the model in three different data sets: one related data set, one independent data set, and one data set of *in vitro* samples to further assess the predictive reliability and stability of the model. The C-index, calibration analysis, ROC analysis, and DCA in the training and validation sets suggested that it is an excellent diagnostic model with good applicability.

Functional analyses of DEGs and IME analyses indicated that neutrophil activation, migration, and degranulation were activated in AS patients. Also, the relative number or proportion of neutrophils was significantly higher in AS patients. Our result is also confirmed by other researchers who have also suggested that the neutrophil-to-lymphocyte ratio be used as an indicator of AS activity (Mercan et al., 2016; Xu et al., 2020; Gökmen et al., 2015). Meanwhile, neutrophil extracellular trap formation and the complement and coagulation cascades were also upregulated in AS, which might induce an autoimmune response, and this is in agreement with the IME analysis result and our current understanding of AS (Gonnet-Gracia et al., 2008; Yang et al., 2016b). A potential explanation for the aforementioned finding is that the increased number of neutrophils might release excessive IL-17A, the key cytokine in the pathogenesis of AS. Although mature neutrophils lack the transcriptional machinery to produce IL-17A, they could produce and store IL-17A before they mature and accumulate it from the extracellular environment (Lin et al., 2011; Tamassia et al., 2018). Besides, in GSE25101, monocytes were also found to be more numerous in AS patients with DEGs clustered in myeloid cell differentiation and leukocyte migration in GO clustering. It is known that monocytes share some similar functions with neutrophils in immune response, and there have also been reports that the monocyte-to-lymphocyte ratio was increased in AS patients (Huang et al., 2018; Wang et al., 2021; Liang et al., 2021). Whether or not the increments in the number of lymphocytes and monocytes are two different subtypes of AS remains unknown.

Lastly, *IL2RB* is a hub gene both in GO/KEGG clustering and the PPI network analysis. Its expression was positively correlated with the relative quantities/proportions of resting NK cells and negatively correlated with the relative quantities/proportions of neutrophils and monocytes in our study, which is in line with the data from the Human Protein Atlas (HPA) website (Karlsson et al., 2021) (Figure 7A: available from v21.1.proteinatlas.org, https://www.proteinatlas.org/ENSG00000100385-IL2RB/single+cell+type). While *ZDHHC18* was observed to be positively correlated with the relative quantities/proportions of neutrophils in all three sets, it did not seem to be highly expressed in granulocytes based on the data from the HPA website (Karlsson et al., 2021) (Figure 7B: available from v21.1.proteinatlas.org, https://www.proteinatlas.org/ENSG00000100385-IL2RB/single+cell+type). Above all, our results suggests that *IL2RB* might be correlated with AS *via* the suppression of the function of resting NK cells, and *ZDHHC18* might be correlated with AS through the function of neutrophils; however, the detailed underlying mechanism still needs to be studied further.

In this study, *IL2RB* and *ZDHHC18* were the two finally screened hub genes. The former had already been identified by other researchers as one of the hub genes in AS (Zhu et al., 2013; Zheng et al., 2021), while to the best of our knowledge, the latter was first reported here by us. *IL2RB*, interleukin 2 receptor subunit beta, encoded the beta subunit of a heterodimer or heterotrimer receptor involved in T cell–mediated immune responses and is probably involved in the stimulation of neutrophil phagocytosis by *IL15* (Ratthé and Girard, 2004;

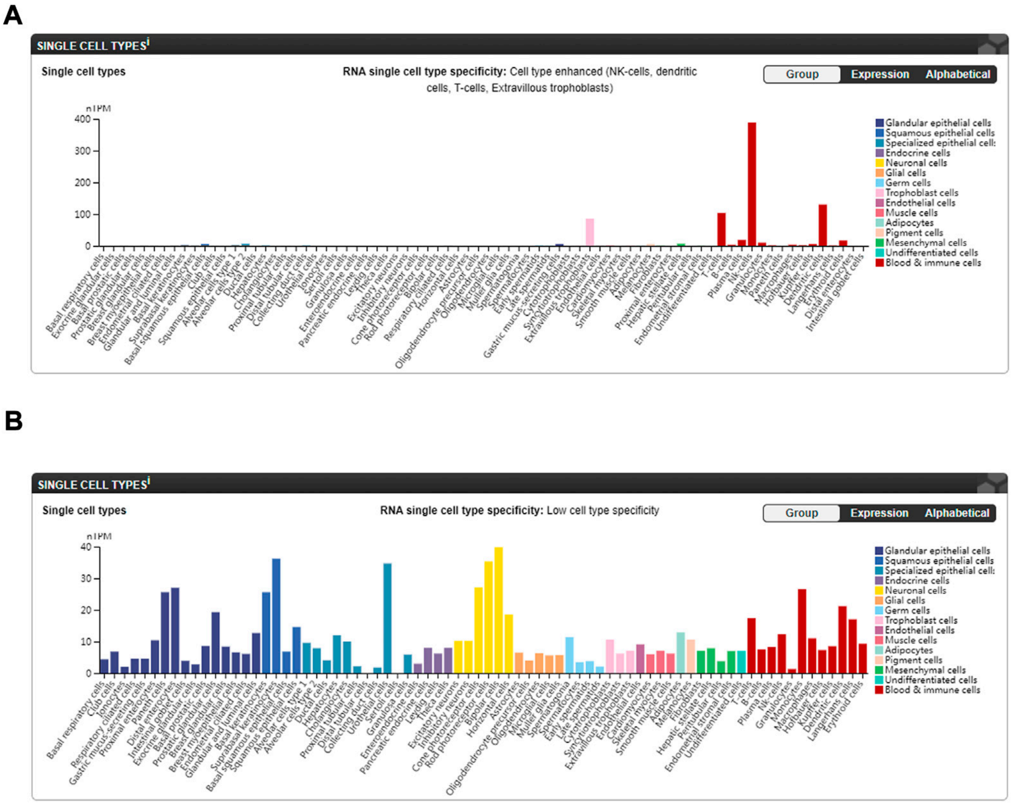


FIGURE 7
Expression of hub genes in different single-cell types of normal subjects from the HPA website (Karlsson et al., 2021). **(A)** Expression of IL2RB in different single-cell types (available from v21.1. proteinatlas.org: https://www.proteinatlas.org/ENSG00000100385-IL2RB/single+cell+type). **(B)** Expression of ZDHHC18 in different single-cell types (available from v21.1. proteinatlas.org: https://www.proteinatlas.org/ENSG00000204160-ZDHHC18/single+cell+type).

Zhang et al., 2019). This protein is a type-I membrane protein primarily expressed in NK cells, T cells, and dendritic cells. According to the KEGG database (https://www.kegg.jp/), *IL2RB* was involved in many pathways, which include endocytosis, the PI3K-Akt signaling pathway, the JAK-STAT signaling pathway, Th1 and Th2 cell differentiation, Th17 cell differentiation, and many more. Obviously, Th1 and Th2 cell differentiation and Th17 cell differentiation seemed to be most related to AS, such that IL2 signaling can inhibit the differentiation of Th17 *via* the inhibition of the transcription factor *RORγt* (Waldmann, 2006; Soper et al., 2007; Liao et al., 2011; Allard-Chamard et al., 2020; Pol et al., 2020). Therefore, with the downregulation of *IL2RB* in this study, Th17 was anticipated to be expanded. However, Th1 and Th2 cell differentiation and Th17 cell differentiation were observed to be downregulated in GSEA (Figure 6D), which is contradictory to our knowledge of AS; therefore, something should be noticed. On the one hand, the pathogeneses of changes in AS are mainly involved in the focus of AS, not in the circulatory system, and our knowledge was largely based on that; so, it might be common for samples from the two sites to have some differences. On the other hand, the role of *IL2* signaling in the differentiation of Th17 has still not been fully clarified (Campbell and Bryceson, 2019). The question is what was the minimum *IL2* signal required to maintain the Treg numbers. Isabel Z Fernandez et al. reported a hypomorphic mutation of *IL2RB* in two infant siblings that resulted in an anticipated reduction in Tregs and an expansion of immature NK cells (Fernandez et al., 2019). Here, in two of the three sets, the relative numbers of $CD8^+$ and gamma-delta T cells were decreased, while that of Tregs was not significantly reduced, which might indicate that the reduced *IL2* signal was still adequate for the proliferation of Tregs and the suppression of effector T-cell expansion (Figures 5D,F). Besides, *via* the blockade of *IL-2 in vitro* and *in vivo*, Kenjiro Fujimura et al. found that the number of Th17 cells did not significantly increase but the proportion of Th17 cells did, which suggests that it might increase the proportion of Th17 by suppressing the total number of immune cells (Fujimura et al., 2013). In this study, the numbers of certain kinds of T cells, such as $CD8^+$ T cells gamma-delta T cells, and Tregs, were observed to have decreased in AS patients, and this might overwhelm the effect of the downregulation of Th17 cell differentiation. However, to see which type of immune cells became fewer and if this would affect the synthesis of IL17 by Th17 cells in AS patients requires further research. In the end, although the potential function of *IL2RB* in AS remains unclear, it might contribute to AS by reducing the number of Treg cells and relatively increasing the proportion of Th17 cells, thereby activating the *IL17* signaling to form AS phenotypes.

ZDHHC18, zinc finger DHHC-type palmitoyltransferase 18, encoded a palmitoyltransferase, which was involved in peptidyl-L-cysteine S-palmitoylation (Ohno et al., 2012). Studies on *ZDHHC18* are rare and insufficient. Currently, it is reported to be associated with innate immunity (Shi et al., 2022), glioma (Chen et al., 2019), ovarian cancer (Pei et al., 2022), and schizophrenia (Zhao et al., 2018). The common palmitoylation substrates of ZDHHC18 are HRAS and LCK (Baumgart et al., 2010; Akimzhanov and Boehning, 2015; Adachi et al., 2016). Palmitoylated HRAS could be translocated and stably anchored to the plasma membrane (Yang et al., 2020), while palmitoylation-defective HRAS was trapped in the Golgi apparatus and was unable to traverse to the plasma membrane. Meanwhile, *ZDHHC18* could activate the rap1 signaling pathway by the palmitoylation of Ras and promote the proliferation of cells, which was consistent with our GSEA result. Besides, Rac1, which was also involved in the rap1 signaling pathway mainly by regulating cell adhesion, migration, and polarity, could also be palmitoylated by the ZDHHC family (Yang et al., 2020). Though we currently do not know the exact role of *ZDHHC18* in AS, it is essential for neutrophil motility as well as directional sensing during migration, which was clustered by GO clustering in our study. In addition, the palmitoylation of LCK could promote T-cell receptor signaling to activate T cells, although this was not seen in our study, which meant that it is not important in the pathogenesis of AS. Furthermore, *ZDHHC18* could negatively regulate CGAS-STING signaling–mediated antiviral innate immunity *via* the palmitoylation of cGAS, which means that the antiviral immunity in AS patients might be impaired by the high expression of *ZDHHC18* (Shi et al., 2022). In our study, KEGG and GSEA also indicated dysregulation in some antiviral immune pathways.

In general, our study indicated that *IL2RB* might be involved in the pathogenesis of AS through the *IL2* signaling pathway and *ZDHHC18* through the rap1 signaling pathway. Both of these could be used as potential biomarkers in AS. Meanwhile, it should also be noted that although we explored some changes in RNA expression in the peripheral blood of AS patients, it is only just the tip of the iceberg. Therefore, more validations of the two genes in AS patients are required, and the mechanisms of these two genes in the pathogenesis of AS also require further research. These are the two main directions of our subsequent research.

Conclusion

IL2RB and *ZDHHC18* were identified as potential blood biomarkers of AS, which might be used for the early diagnosis of AS and serve as supplements to the existing diagnostic methods. Our study helps deepen the understanding of the pathogenesis of AS.

Data availability statement

Publicly available data sets were analyzed in this study. These data can be found at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73754 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE25101 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18781 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11886.

Author contributions

JW and LW conceived the study. JW performed the analyses by R and wrote the draft. XD reviewed the manuscript. All authors contributed to the manuscript and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, editors, and reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adachi, N., Hess, D. T., McLaughlin, P., and Stamler, J. S. (2016). S-palmitoylation of a novel site in the β2-adrenergic receptor associated with a novel intracellular itinerary. *J. Biol. Chem.* 291 (38), 20232–20246. doi:10.1074/jbc.M116.725762

Akimzhanov, A. M., and Boehning, D. (2015). Rapid and transient palmitoylation of the tyrosine kinase Lck mediates Fas signaling. *Proc. Natl. Acad. Sci. U. S. A.* 112 (38), 11876–11880. doi:10.1073/pnas.1509929112

Allard-Chamard, H., Mishra, H. K., Nandi, M., Mayhue, M., Menendez, A., Ilangumaran, S., et al. (2020). Interleukin-15 in autoimmunity. *Cytokine* 136, 155258. doi:10.1016/j.cyto.2020.155258

Baumgart, F., Corral-Escariz, M., Pérez-Gil, J., and Rodríguez-Crespo, I. (2010). Palmitoylation of R-Ras by human DHHC19, a palmitoyl transferase with a CaaX box. *Biochim. Biophys. Acta* 1798 (3), 592–604. doi:10.1016/j.bbamem.2010.01.002

Bowness, P. (2015). HLA-B27. *Annu. Rev. Immunol.* 33, 29–48. doi:10.1146/annurev-immunol-032414-112110

Bowness, P., Ridley, A., Shaw, J., Chan, A. T., Wong-Baeza, I., Fleming, M., et al. (2011). Th17 cells expressing KIR3DL2+ and responsive to HLA-B27 homodimers are increased in ankylosing spondylitis. *J. Immunol.* 186 (4), 2672–2680. doi:10.4049/jimmunol.1002653

Campbell, T. M., and Bryceson, Y. T. (2019). IL2RB maintains immune harmony. *J. Exp. Med.* 216 (6), 1231–1233. doi:10.1084/jem.20190546

Chen, X., Hu, L., Yang, H., Ma, H., Ye, K., Zhao, C., et al. (2019). DHHC protein family targets different subsets of glioma stem cells in specific niches. *J. Exp. Clin. Cancer Res.* 38 (1), 25. doi:10.1186/s13046-019-1033-2

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., and Campbell, J. P. (2020). Introduction to machine learning, neural networks, and deep learning. *Transl. Vis. Sci. Technol.* 9 (2), 14.

Colbert, R. A., DeLay, M. L., Klenk, E. I., and Layh-Schmitt, G. (2010). From HLA-B27 to spondyloarthritis: A journey through the ER. *Immunol. Rev.* 233 (1), 181–202. doi:10.1111/j.0105-2896.2009.00865.x

Danve, A., and O'Dell, J. (2015). The ongoing quest for biomarkers in ankylosing spondylitis. *Int. J. Rheum. Dis.* 18 (8), 826–834.

Dean, L. E., Jones, G. T., MacDonald, A. G., Downham, C., Sturrock, R. D., and Macfarlane, G. J. (2014). Global prevalence of ankylosing spondylitis. *Rheumatol. Oxf.* 53 (4), 650–657. doi:10.1093/rheumatology/ket387

Fernandez, I. Z., Baxter, R. M., Garcia-Perez, J. E., Vendrame, E., Ranganath, T., Kong, D. S., et al. (2019). A novel human IL2RB mutation results in T and NK cell-driven immune dysregulation. *J. Exp. Med.* 216 (6), 1255–1267. doi:10.1084/jem.20182015

Fujimura, K., Oyamada, A., Iwamoto, Y., Yoshikai, Y., and Yamada, H. (2013). CD4 T cell-intrinsic IL-2 signaling differentially affects Th1 and Th17 development. *J. Leukoc. Biol.* 94 (2), 271–279. doi:10.1189/jlb.1112581

Gökmen, F., Akbal, A., Reşorlu, H., Gökmen, E., Güven, M., Aras, A. B., et al. (2015). Neutrophil-lymphocyte ratio connected to treatment options and inflammation markers of ankylosing spondylitis. *J. Clin. Lab. Anal.* 29 (4), 294–298. doi:10.1002/jcla.21768

Gonnet-Gracia, C., Barnetche, T., Richez, C., Blanco, P., Dehais, J., and Schaeverbeke, T. (2008). Anti-nuclear antibodies, anti-DNA and C4 complement evolution in rheumatoid arthritis and ankylosing spondylitis treated with TNF-alpha blockers. *Clin. Exp. Rheumatol.* 26 (3), 401–407.

Guiliano, D. B., North, H., Panayoitou, E., Campbell, E. C., McHugh, K., Cooke, F. G., et al. (2017). Polymorphisms in the F pocket of HLA-B27 subtypes strongly affect assembly, chaperone interactions, and heavy-chain misfolding. *Arthritis Rheumatol.* 69 (3), 610–621. doi:10.1002/art.39948

Huang, Y., Deng, W., Zheng, S., Feng, F., Huang, Z., Huang, Q., et al. (2018). Relationship between monocytes to lymphocytes ratio and axial spondyloarthritis. *Int. Immunopharmacol.* 57, 43–46. doi:10.1016/j.intimp.2018.02.008

Hu, Q., Sun, Y., Li, Y., Shi, H., Teng, J., Liu, H., et al. (2018). Anti-SIRT1 autoantibody is elevated in ankylosing spondylitis: A potential disease biomarker. *BMC Immunol.* 19 (1), 38.

Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., et al. (2021). A single-cell type transcriptomics map of human tissues. *Sci. Adv.* 7 (31), eabh2169. doi:10.1126/sciadv.abh2169

Kook, H. Y., Jin, S. H., Park, P. R., Lee, S. J., Shin, H. J., and Kim, T. J. (2019). Serum miR-214 as a novel biomarker for ankylosing spondylitis. *Int. J. Rheum. Dis.* 22 (7), 1196–1201.

Liang, T., Chen, J., Xu, G., Zhang, Z., Xue, J., Zeng, H., et al. (2021). Platelet-to-Lymphocyte ratio as an independent factor was associated with the severity of ankylosing spondylitis. *Front. Immunol.* 12, 760214. doi:10.3389/fimmu.2021.760214

Liao, W., Lin, J. X., Wang, L., Li, P., and Leonard, W. J. (2011). Modulation of cytokine receptors by IL-2 broadly regulates differentiation into helper T cell lineages. *Nat. Immunol.* 12 (6), 551–559. doi:10.1038/ni.2030

Lin, A. M., Rubin, C. J., Khandpur, R., Wang, J. Y., Riblett, M., Yalavarthi, S., et al. (2011). Mast cells and neutrophils release IL-17 through extracellular trap formation in psoriasis. *J. Immunol.* 187 (1), 490–500. doi:10.4049/jimmunol.1100123

Mauro, D., Thomas, R., Guggino, G., Lories, R., Brown, M. A., and Ciccia, F. (2021). Ankylosing spondylitis: An autoimmune or autoinflammatory disease? *Nat. Rev. Rheumatol.* 17 (7), 387–404. doi:10.1038/s41584-021-00625-y

Mear, J. P., Schreiber, K. L., Münz, C., Zhu, X., Stevanović, S., Rammensee, H. G., et al. (1999). Misfolding of HLA-B27 as a result of its B pocket suggests a novel mechanism for its role in susceptibility to spondyloarthropathies. *J. Immunol.* 163 (12), 6665–6670.

Mercan, R., Bitik, B., Tufan, A., Bozbulut, U. B., Atas, N., Ozturk, M. A., et al. (2016). The association between neutrophil/lymphocyte ratio and disease activity in rheumatoid arthritis and ankylosing spondylitis. *J. Clin. Lab. Anal.* 30 (5), 597–601. doi:10.1002/jcla.21908

Navarro-Compán, V., Sepriano, A., El-Zorkany, B., and van der Heijde, D. (2021). Axial spondyloarthritis. *Ann. Rheum. Dis.* 80 (12), 1511–1521. doi:10.1136/annrheumdis-2021-221035

Ohno, Y., Kashio, A., Ogata, R., Ishitomi, A., Yamazaki, Y., and Kihara, A. (2012). Analysis of substrate specificity of human DHHC protein acyltransferases using a yeast expression system. *Mol. Biol. Cell.* 23 (23), 4543–4551. doi:10.1091/mbc.E12-05-0336

Pedersen, S. J., and Maksymowych, W. P. (2019). The pathogenesis of ankylosing spondylitis: An update. *Curr. Rheumatol. Rep.* 21 (10), 58. doi:10.1007/s11926-019-0856-3

Pei, X., Li, K. Y., Shen, Y., Li, J. T., Lei, M. Z., Fang, C. Y., et al. (2022). Palmitoylation of MDH2 by ZDHHC18 activates mitochondrial respiration and accelerates ovarian cancer growth. *Sci. China. Life Sci.* 65, 2017–2030. doi:10.1007/s11427-021-2048-2

Perrotta, F. M., Ceccarelli, F., Barbati, C., Colasanti, T., De Socio, A., Scriffignano, S., et al. (2022). Serum sclerostin as a possible biomarker in ankylosing spondylitis: A case-control study. *J. Immunol. Res.* 2018, 9101964.

Pol, J. G., Caudana, P., Paillet, J., Piaggio, E., and Kroemer, G. (2020). Effects of interleukin-2 in immunostimulation and immunosuppression. *J. Exp. Med.* 217 (1), e20191247. doi:10.1084/jem.20191247

Ratthé, C., and Girard, D. (2004). Interleukin-15 enhances human neutrophil phagocytosis by a syk-dependent mechanism: Importance of the IL-15Ralpha chain. *J. Leukoc. Biol.* 76 (1), 162–168. doi:10.1189/jlb.0605298

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47. doi:10.1093/nar/gkv007

Ritchlin, C., and Adamopoulos, I. E. (2021). Axial spondyloarthritis: New advances in diagnosis and management. *Bmj* 372, m4447. doi:10.1136/bmj.m4447

Rudwaleit, M., van der Heijde, D., Landewé, R., Akkoc, N., Brandt, J., Chou, C. T., et al. (2011). The Assessment of SpondyloArthritis International Society classification criteria for peripheral spondyloarthritis and for spondyloarthritis in general. *Ann. Rheum. Dis.* 70 (1), 25–31. doi:10.1136/ard.2010.133645

Rudwaleit, M., van der Heijde, D., Landewé, R., Listing, J., Akkoc, N., Brandt, J., et al. (2009). The development of assessment of SpondyloArthritis international society classification criteria for axial spondyloarthritis (part II): Validation and final selection. *Ann. Rheum. Dis.* 68 (6), 777–783. doi:10.1136/ard.2009.108233

Sanz, H., Valim, C., Vegas, E., Oller, J. M., and Reverter, F. (2018). SVM-RFE: Selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinforma.* 19 (1), 432. doi:10.1186/s12859-018-2451-4

Sharip, A., and Kunz, J. (2020). Understanding the pathogenesis of spondyloarthritis. *Biomolecules* 10 (10), E1461. doi:10.3390/biom10101461

Shi, C., Yang, X., Liu, Y., Li, H., Chu, H., Li, G., et al. (2022). ZDHHC18 negatively regulates cGAS-mediated innate immunity through palmitoylation. *Embo J.* 41, e109272. doi:10.15252/embj.2021109272

Sieper, J., Braun, J., Dougados, M., and Baeten, D. (2015). Axial spondyloarthritis. *Nat. Rev. Dis. Prim.* 1, 15013. doi:10.1038/nrdp.2015.13

Sieper, J., and Poddubnyy, D. (2017). Axial spondyloarthritis. *Lancet* 390 (10089), 73–84. doi:10.1016/S0140-6736(16)31591-4

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* 39 (5), 1–13. doi:10.18637/jss.v039.i05

Soper, D. M., Kasprowicz, D. J., and Ziegler, S. F. (2007). IL-2Rbeta links IL-2R signaling with Foxp3 expression. *Eur. J. Immunol.* 37 (7), 1817–1826. doi:10.1002/eji.200737101

Stolwijk, C., van Onna, M., Boonen, A., and van Tubergen, A. (2016). Global prevalence of spondyloarthritis: A systematic review and meta-regression analysis. *Arthritis Care Res. Hob.* 68 (9), 1320–1331. doi:10.1002/acr.22831

Strobl, C., Boulesteix, A. L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinforma.* 8, 25. doi:10.1186/1471-2105-8-25

Sun, Y., Ouyang, B., Xie, Q., Wang, L., Zhu, S., and Jia, Y. (2020). Serum Deoxyribonuclease 1-like 3 is a potential biomarker for diagnosis of ankylosing spondylitis. *Clin. Chim. Acta.* 503, 197–202.

Tamassia, N., Arruda-Silva, F., Calzetti, F., Lonardi, S., Gasperini, S., Gardiman, E., et al. (2018). A reappraisal on the potential ability of human neutrophils to express and produce IL-17 family members *in vitro*: Failure to reproducibly detect it. *Front. Immunol.* 9, 795. doi:10.3389/fimmu.2018.00795

Taurog, J. D., Chhabra, A., and Colbert, R. A. (2016). Ankylosing spondylitis and axial spondyloarthritis. *N. Engl. J. Med.* 374 (26), 1303–1374. doi:10.1056/NEJMc1609622

van der Linden, S., Valkenburg, H. A., and Cats, A. (1984). Evaluation of diagnostic criteria for ankylosing spondylitis. A proposal for modification of the New York criteria. *Arthritis Rheum.* 27 (4), 361–368. doi:10.1002/art.1780270401

Voruganti, A., and Bowness, P. (2020). New developments in our understanding of ankylosing spondylitis pathogenesis. *Immunology* 161 (2), 94–102. doi:10.1111/imm.13242

Waldmann, T. A. (2006). The biology of interleukin-2 and interleukin-15: Implications for cancer therapy and vaccine design. *Nat. Rev. Immunol.* 6 (8), 595–601. doi:10.1038/nri1901

Wang, J., Su, J., Yuan, Y., Jin, X., Shen, B., and Lu, G. (2021). The role of lymphocyte-monocyte ratio on axial spondyloarthritis diagnosis and sacroiliitis staging. *BMC Musculoskelet. Disord.* 22 (1), 86. doi:10.1186/s12891-021-03973-8

Ward, M. M., Deodhar, A., Gensler, L. S., Dubreuil, M., Yu, D., Khan, M. A., et al. (2019). 2019 update of the American College of rheumatology/spondylitis association of America/spondyloarthritis research and treatment network recommendations for the treatment of ankylosing spondylitis and nonradiographic axial spondyloarthritis. *Arthritis Rheumatol.* 71 (10), 1599–1613. doi:10.1002/art.41042

Xu, S., Ma, Y., Wu, M., Zhang, X., Yang, J., Deng, J., et al. (2020). Neutrophil lymphocyte ratio in patients with ankylosing spondylitis: A systematic review and meta-analysis. *Mod. Rheumatol.* 30 (1), 141–148. doi:10.1080/14397595.2018.1564165

Yang, C., Ding, P., Wang, Q., Zhang, L., Zhang, X., Zhao, J., et al. (2016b). Inhibition of complement retards ankylosing spondylitis progression. *Sci. Rep.* 6, 34643. doi:10.1038/srep34643

Yang, L., Wang, L., Wang, X., Xian, C. J., and Lu, H. (2016a). A possible role of intestinal microbiota in the pathogenesis of ankylosing spondylitis. *Int. J. Mol. Sci.* 17 (12), E2126. doi:10.3390/ijms17122126

Yang, X., Chatterjee, V., Ma, Y., Zheng, E., and Yuan, S. Y. (2020). Protein palmitoylation in leukocyte signaling and function. *Front. Cell. Dev. Biol.* 8, 600368. doi:10.3389/fcell.2020.600368

Yu, H. C., Huang, K. Y., Lu, M. C., Huang, H. L., Liu, S. Q., Lai, N. S., et al. (2017). Targeted delivery of the HLA-B(*)27-Binding peptide into the endoplasmic reticulum suppresses the IL-23/IL-17 Axis of immune cells in spondylarthritis. *Mediat. Inflamm.* 2017, 4016802. doi:10.1155/2017/4016802

Zhang, Z., Gothe, F., Pennamen, P., James, J. R., McDonald, D., Mata, C. P., et al. (2019). Human interleukin-2 receptor β mutations associated with defects in immunity and peripheral tolerance. *J. Exp. Med.* 216 (6), 1311–1327. doi:10.1084/jem.20182304

Zhao, Y., He, A., Zhu, F., Ding, M., Hao, J., Fan, Q., et al. (2018). Integrating genome-wide association study and expression quantitative trait locus study identifies multiple genes and gene sets associated with schizophrenia. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 81, 50–54. doi:10.1016/j.pnpbp.2017.10.003

Zheng, Y., Cai, B., Ren, C., Xu, H., Du, W., Wu, Y., et al. (2021). Identification of immune related cells and crucial genes in the peripheral blood of ankylosing spondylitis by integrated bioinformatics analysis. *PeerJ* 9, e12125. doi:10.7717/peerj.12125

Zhu, Z. Q., Tang, J. S., and Cao, X. J. (2013). Transcriptome network analysis reveals potential candidate genes for ankylosing spondylitis. *Eur. Rev. Med. Pharmacol. Sci.* 17 (23), 3178–3185.

Zochling, J., Brandt, J., and Braun, J. (2005). The current concept of spondyloarthritis with special emphasis on undifferentiated spondyloarthritis. *Rheumatol. Oxf.* 44 (12), 1483–1491. doi:10.1093/rheumatology/kei047

TYPE Original Research
PUBLISHED 14 February 2023
DOI 10.3389/fgene.2023.1095330

3PNMF-MKL: A non-negative matrix factorization-based multiple kernel learning method for multi-modal data integration and its application to gene signature detection

Saurav Mallik[1]*, Anasua Sarkar[2], Sagnik Nath[2], Ujjwal Maulik[2], Supantha Das[3], Soumen Kumar Pati[4], Soumadip Ghosh[5] and Zhongming Zhao[6,7]*

[1]Department of Environmental Health, Harvard T H Chan School of public Health, Boston, MA, United States, [2]Department of Computer Science & Engineering, Jadavpur University, Kolkata, India, [3]Department of Information Technology, Academy of Technology, Hooghly, West Bengal, India, [4]Department of Bioinformatics, Maulana Abul Kalam Azad University, Kolkata, West Bengal, India, [5]Department of Computer Science & Engineering, Sister Nivedita University, New Town, West Bengal, India, [6]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States, [7]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States

Check for updates

OPEN ACCESS

EDITED BY
Andrei Rodin,
City of Hope National Medical Center,
United States

REVIEWED BY
Chunhou Zheng,
Anhui University, China
Loveleen Gaur,
Amity University, India

*CORRESPONDENCE
Saurav Mallik,
✉ sauravmtech2@gmail.com,
✉ smallik@hsph.harvard.edu
Zhongming Zhao,
✉ zhongming.zhao@uth.tmc.edu

SPECIALTY SECTION
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

RECEIVED 11 November 2022
ACCEPTED 30 January 2023
PUBLISHED 14 February 2023

CITATION
Mallik S, Sarkar A, Nath S, Maulik U, Das S, Pati SK, Ghosh S and Zhao Z (2023), 3PNMF-MKL: A non-negative matrix factorization-based multiple kernel learning method for multi-modal data integration and its application to gene signature detection.
Front. Genet. 14:1095330.
doi: 10.3389/fgene.2023.1095330

In this current era, biomedical big data handling is a challenging task. Interestingly, the integration of multi-modal data, followed by significant feature mining (gene signature detection), becomes a daunting task. Remembering this, here, we proposed a novel framework, namely, three-factor penalized, non-negative matrix factorization-based multiple kernel learning with soft margin hinge loss (3PNMF-MKL) for multi-modal data integration, followed by gene signature detection. In brief, limma, employing the empirical Bayes statistics, was initially applied to each individual molecular profile, and the statistically significant features were extracted, which was followed by the three-factor penalized non-negative matrix factorization method used for data/matrix fusion using the reduced feature sets. Multiple kernel learning models with soft margin hinge loss had been deployed to estimate average accuracy scores and the area under the curve (AUC). Gene modules had been identified by the consecutive analysis of average linkage clustering and dynamic tree cut. The best module containing the highest correlation was considered the potential gene signature. We utilized an acute myeloid leukemia cancer dataset from The Cancer Genome Atlas (TCGA) repository containing five molecular profiles. Our algorithm generated a 50-gene signature that achieved a high classification AUC score (viz., 0.827). We explored the functions of signature genes using pathway and Gene Ontology (GO) databases. Our method outperformed the state-of-the-art methods in terms of computing AUC. Furthermore, we included some comparative studies with other related methods to enhance the acceptability of our method. Finally, it can be notified that our algorithm can be applied to any multi-modal dataset for data integration, followed by gene module discovery.

KEYWORDS

multi-omics, gene signature detection, feature selection, DNA methylation, matrix factorization

1 Introduction

Rapid advances in biotechnology have enabled the generation of data in multiple platforms from the same or similar bio-samples. For example, The Cancer Genome Atlas (TCGA) comprehensively generated multi-omics profiles in 33 cancer types and subtypes. Therefore, it is made available to conduct an in-depth investigation into various molecular incidents at different biological stages and for specific tumor categories. The challenging task here is to develop algorithms to properly integrate these multi-omics (i.e., multi-modal) data, which will deepen our understanding of human tumorigenesis.

The integration of multi-omics profiles is a fast emerging area of the biomedical research (Imielinski et al., 2012; Mo et al., 2013; Mallik et al., 2017; Gaur et al., 2022; Ghose et al., 2022; Saeed et al., 2022). From the perspective of biology, cellular processes are based on the communication among different biomolecules (viz., mutations, epigenetic regulators, proteins, and metabolites). Molecular regulations occur in multi-layers and multi-vantage points to orchestrate complex biological events. An integrated analysis of profiles on the common set of samples from multi-omics data shows great potential to yield more biologically meaningful outcomes over an individual analysis on a single data layer. Overall, it shows a more comprehensive view and a global functional orientation of the biological system.

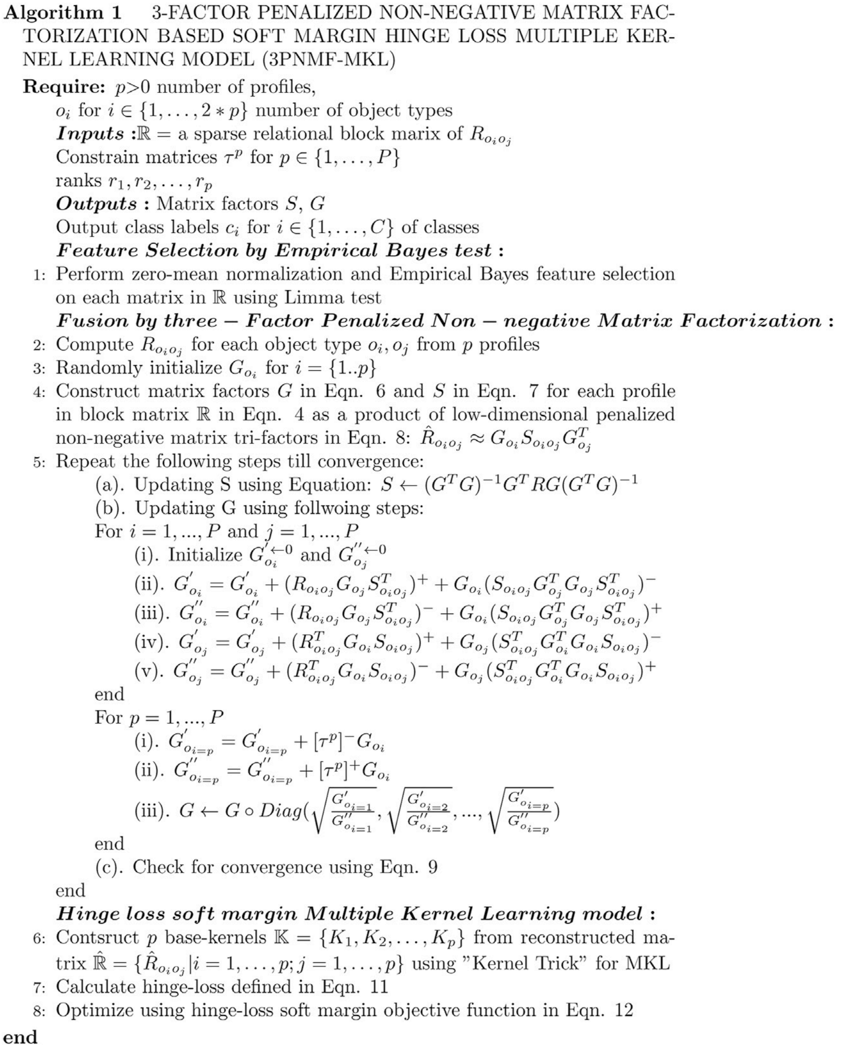
One of the major challenges for integration is to deal with the heterogeneity of these profiles. Profiles from various sources are often complicated to integrate or interpret together because of the inherent discrepancies. Various genomic variables can be measured and accumulated in different ways, which are also vulnerable to different kinds of noise and various confounding effects. Interestingly, these profiles show individual aspects of the biological system at different angles. The discrepancy among multi-omics data, therefore, provides an opportunity for detecting reliable and consistent signals for biological studies in a comprehensive manner. Multi-dimensional data integration and gene signature identification are among the most challenging tasks for bioinformaticians (Li et al., 2019; Mallik and Zhao, 2020; Qiu et al., 2020; Pellet et al., 2015; Serra et al., 2015). Mallik et al. (2017) proposed a scheme to recognize epigenetic biomarkers applying maximal relevance and minimal redundancy-based feature selection for multi-omics data. An approach of the integration of multi-omics data was proposed by Li et al. (2019) to identify biomarkers in the domain of cancer research. Qiu et al. (2020) suggested an approach regarding the revelation of 172 osteoporosis biomarkers by multi-omics data integration. A scheme of multi-omics data integration was presented by Pellet et al. (2015) to determine predictive molecular signatures regarding CLAD. Because specific profiles contain different characteristics/phenomena, integration of multi-view data with significant feature reduction and gene signature detection is fundamentally important. In this upcoming era, the multi-platform integration approach has been applied to accomplish various important tasks, such as signature/bio-marker detection, disease classification, and gene clustering. Prior research works in bio-marker discovery (Bandyopadhyay and Mallik, 2016; Kandimalla et al., 2022), classification (Henry et al., 2014; Maulik et al., 2015; Zhang and Kuster, 2019), and clustering (Wang and Gu, 2016) have improved the promising performance of multi-modal integration approaches. Nevertheless, the outcomes of such approaches are not always satisfactory. Zhang and Kuster (2019) represented an approach with the incorporation of proteomics data to express the significance of omics data integration with higher accuracy. Kandimalla et al. (2022) showed mRNA–miRNA regulatory network analyses to improve the approach of multi-omics data integration. In this work, we propose a novel framework, namely three-factor penalized non-negative matrix factorization-based multiple kernel learning with soft margin hinge loss (3PNMF-MKL), which applies consecutive utilization of a couple of multi-dimensional strategies: i) statistical empirical Bayes-based feature selection, ii) three-factor penalized non-negative matrix factorization, iii) multiple kernel learning with soft margin hinge loss, iv) average linkage clustering, and v) the dynamic tree cut method for multi-platform data integration and gene signature detection. For evaluation of the performance of our proposed approach, a cancer dataset from TCGA acute myeloid leukemia (LAML) containing five different profiles [gene expression, DNA methylation, exon expression, pathway activity, and copy number variation (CNV)] was used. We demonstrated that our approach is capable of multi-modal data integration, and thus, it can be applied to any kind of multi-platform datasets.

2 Experimental procedures

In this section, we illustrate our proposed approach for identifying Pareto-optimal gene signatures by feature clustering on a cancer multi-omics dataset. The major steps are described as follows.

2.1 Feature selection by the empirical Bayes test

Commonly shared features (genes/probes) and samples are chosen across all the profiles from the multi-omics cancer dataset. Specifically, probes (features) from DNA methylation arrays containing any missing values are discarded. The individual profile is normalized using the zero-mean normalization for each feature (Bandyopadhyay et al., 2013), as described in the following formula: $x'_{ik} = \frac{x_{ik}-\mu}{\sigma}$. Here, μ is the mean across the data for the feature i prior to normalization, and σ denotes standard deviation. x_{ik} and x'_{ik} signify the value of the i-th feature at k-th patient (sample) prior and after normalization, respectively. To determine statistically significant features, the empirical Bayes statistical test is applied using the package "Linear Models for Microarray and RNA-Seq Data" (Smyth, 2004; Bandyopadhyay et al., 2013), which works better on the dataset with a small sample size. The moderated t-statistic (Ritchie et al., 2015) is elaborated as follows:



```
Algorithm 1    3-FACTOR PENALIZED NON-NEGATIVE MATRIX FACTORIZATION BASED SOFT MARGIN HINGE LOSS MULTIPLE KERNEL LEARNING MODEL (3PNMF-MKL)
  Require: $p$>0 number of profiles,
    $o_i$ for $i \in \{1, \dots, 2 * p\}$ number of object types
    Inputs :$\mathbb{R}$ = a sparse relational block marix of $R_{o_i o_j}$
    Constrain matrices $\tau^p$ for $p \in \{1, \dots, P\}$
    ranks $r_1, r_2, \dots, r_p$
    Outputs : Matrix factors $S$, $G$
    Output class labels $c_i$ for $i \in \{1, \dots, C\}$ of classes
    Feature Selection by Empirical Bayes test :
    Perform zero-mean normalization and Empirical Bayes feature selection
    on each matrix in $\mathbb{R}$ using Limma test
    Fusion by three − Factor Penalized Non − negative Matrix Factorization :
    Compute $R_{o_i o_j}$ for each object type $o_i, o_j$ from $p$ profiles
    Randomly initialize $G_{o_i}$ for $i = \{1..p\}$
    Construct matrix factors $G$ in Eqn. 6 and $S$ in Eqn. 7 for each profile
    in block matrix $\mathbb{R}$ in Eqn. 4 as a product of low-dimensional penalized
    non-negative matrix tri-factors in Eqn. 8: $\hat{R}_{o_i o_j} \approx G_{o_i} S_{o_i o_j} G_{o_j}^T$
    Repeat the following steps till convergence:
        (a). Updating S using Equation: $S \leftarrow (G^T G)^{-1} G^T R G (G^T G)^{-1}$
        (b). Updating G using follwoing steps:
        For $i = 1, ..., P$ and $j = 1, ..., P$
            (i). Initialize $G_{o_i}^{' \leftarrow 0}$ and $G_{o_j}^{'' \leftarrow 0}$
            (ii). $G_{o_i}^{'} = G_{o_i}^{'} + (R_{o_i o_j} G_{o_j} S_{o_i o_j}^T)^+ + G_{o_i} (S_{o_i o_j} G_{o_j}^T G_{o_j} S_{o_i o_j}^T)^-$
            (iii). $G_{o_i}^{''} = G_{o_i}^{''} + (R_{o_i o_j} G_{o_j} S_{o_i o_j}^T)^- + G_{o_i} (S_{o_i o_j} G_{o_j}^T G_{o_j} S_{o_i o_j}^T)^+$
            (iv). $G_{o_j}^{'} = G_{o_j}^{'} + (R_{o_i o_j}^T G_{o_i} S_{o_i o_j})^+ + G_{o_j} (S_{o_i o_j}^T G_{o_i}^T G_{o_i} S_{o_i o_j})^-$
            (v). $G_{o_j}^{''} = G_{o_j}^{''} + (R_{o_i o_j}^T G_{o_i} S_{o_i o_j})^- + G_{o_j} (S_{o_i o_j}^T G_{o_i}^T G_{o_i} S_{o_i o_j})^+$
        end
        For $p = 1, ..., P$
            (i). $G_{o_{i=p}}^{'} = G_{o_{i=p}}^{'} + [\tau^p]^- G_{o_i}$
            (ii). $G_{o_{i=p}}^{''} = G_{o_{i=p}}^{''} + [\tau^p]^+ G_{o_i}$
            (iii). $G \leftarrow G \circ Diag(\sqrt{\frac{G_{o_{i=1}}'}{G_{o_{i=1}}''}}, \sqrt{\frac{G_{o_{i=2}}'}{G_{o_{i=2}}''}}, ..., \sqrt{\frac{G_{o_{i=p}}'}{G_{o_{i=p}}''}})$
        end
        (c). Check for convergence using Eqn. 9
    end
    Hinge loss soft margin Multiple Kernel Learning model :
    Contsruct $p$ base-kernels $\mathbb{K} = \{K_1, K_2, \dots, K_p\}$ from reconstructed ma-
    trix $\hat{\mathbb{R}} = \{\hat{R}_{o_i o_j} | i = 1, \dots, p; j = 1, \dots, p\}$ using "Kernel Trick" for MKL
    Calculate hinge-loss defined in Eqn. 11
    Optimize using hinge-loss soft margin objective function in Eqn. 12
end
```

FIGURE 1
Algorithm of the proposed 3PNMF-MK model.

$$\tilde{t}_{pr} = \frac{1}{\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \frac{\hat{\beta}_{pr}}{\tilde{s}_{pr}}, \tag{1}$$

where m_1 and m_2 are the number of patients (cases) and that of the normal samples (controls), respectively. Here, $\hat{\beta}_{pr}$ signifies the contrast estimator for the feature *pr*, whereas $\tilde{s}_{pr}^2$ refers to the posterior sample variance for *pr*. The statistic to compute the contrast estimator for the probe *pr* is formulated as follows: $\hat{\beta}_{pr} | \sigma_{pr}^2 \sim N(\beta_{pr}, \sigma_{pr}^2)$. Here, *N* represents the normal distribution. The statistic to estimate the posterior sample variance for *pr* is formulated as follows:

$$\tilde{s}_{pr}^2 = \frac{d_0 s_0^2 + d_{pr} s_{pr}^2}{d_0 + d_{pr}}, \tag{2}$$

where d_0 $(< \infty)$ signifies the prior degrees of freedom, and s_0^2 denotes the variance. In addition, d_{pr} (> 0) symbolizes the experimental degrees of freedom of *pr*, and s_{pr}^2 denotes the sample variance of *pr*. The significance of the level of the *p*-value is then determined from $\tilde{s}_{pr}^2$ with the help of the cumulative distribution function (cdf). If the *p*-value of the feature is less than the standard cutoff of 0.05, the feature is defined as statistically significant. The filtered differentially expressed features are then ordered according to the *p*-values. Notably, if any gene corresponds to more than one probe (feature), the probe with the lowest *p*-value will be selected to represent the gene, and the rest of the probes for the gene are simply ignored. We apply the same approach to each layer of the molecular profile, and then, we perform the combination of the significant non-redundant features (genes/probes/copy number variation, etc.) from all layers (let, *UF*).

2.2 Fusion by matrix factorization

Let o_i and o_j denote two object types, namely, gene expression and DNA methylation, in all resulted features *UF*. The number of genes is N, while each gene is denoted by n_i, where i = 1, 2,. . ., N.

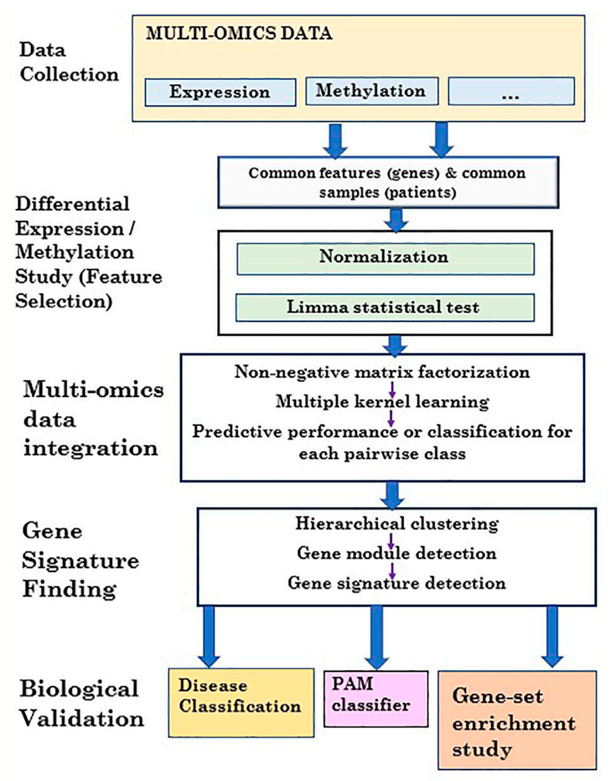

Data Collection
MULTI-OMICS DATA
Expression
Methylation
...
Differential Expression / Methylation Study (Feature Selection)
Common features (genes) & common samples (patients)
Normalization
Limma statistical test
Multi-omics data integration
Non-negative matrix factorization
Multiple kernel learning
Predictive performance or classification for each pairwise class
Gene Signature Finding
Hierarchical clustering
Gene module detection
Gene signature detection
Biological Validation
Disease Classification
PAM classifier
Gene-set enrichment study

FIGURE 2
Flowchart of the proposed 3PNMF-MKL framework.

There are M number of DNA methylation samples, while each sample is termed as m_j, where j = 1, 2, . . ., M. In addition, there is a P set consisting of p types of profiles from the multi-omics datasets. The input to this implemented variant of the 3-*FPNMF* model is $\mathbb{R}$, which is a relational block matrix shown as follows:

$$\mathbb{R} = \begin{bmatrix} * & R_{12} & \dots & R_{1p} \\ R_{21} & * & \dots & R_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p1} & R_{p2} & \dots & * \end{bmatrix}. \tag{3}$$

Here, $*$ denotes that similar object relationships are not considered in this approach. R_{ij} denotes the relationship between o_ith and o_jth object types. The respective correlation of the xth object of type o_i (e.g., gene) and the yth object of type o_j (e.g., sample) is represented as $R_{o_io_j}(x, y)$. In this implementation, we have experimented with six object types, as described later.

For each object type from each profile, there is a constraint in the input constraint block diagonal matrix, as shown in the following expression:

$$\tau^P = Diag(\tau^1, \tau^2, \dots, \tau^P). \tag{4}$$

The relational block matrix $\mathbb{R}$ is tri-factorized into matrix factors G and S (Žitnik and Zupan, 2014), which is shown as follows:

$$G = Diag\left(G^1_{n_1 \times m_1}, G^2_{n_2 \times m_2}, \dots, G^P_{n_p \times m_p}\right), \tag{5}$$

$$S = \begin{bmatrix} * & S_{12}^{r_1 \times r_2} & \dots & S_{1p}^{r_1 \times r_p} \\ S_{21}^{r_2 \times r_1} & * & \dots & S_{2p}^{r_2 \times r_p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1}^{r_p \times r_1} & S_{p2}^{r_p \times r_2} & \dots & * \end{bmatrix}. \tag{6}$$

Here, r denotes rank factorization to the object type o_p inferenced by the 3-*FPNMF* model. The factor S denotes the block relation between object types o_i and o_j. The factor G_{o_i} reconstructs relations specifically to the object type o_i.

Thus, each relation matrix $R_{o_io_j}$ obtains matrix factorization as $G_{o_i}S_{o_io_j}G^T_{o_j}$. In a simplified way, this relational block 3-*FPNMF* model is shown as follows:

$$\begin{bmatrix} * & G_{o_1}S_{o_1o_2}G^T_{o_2} & \dots & G_{o_1}S_{o_1o_p}G^T_{o_p} \\ G_{o_2}S_{o_2o_1}G^T_{o_1} & * & \dots & G_{o_2}S_{o_2o_p}G^T_{o_p} \\ \vdots & \vdots & \ddots & \vdots \\ G_{o_p}S_{o_po_1}G^T_{o_1} & G_{o_p}S_{o_po_2}G^T_{o_2} & \dots & * \end{bmatrix}. \tag{7}$$

The objective function of this tri-factor penalized matrix decomposition (*PMD*) model is to minimize the distance between the input block relational matrix $\mathbb{R}$ and its 3-*FPNMF* system adhering to the constraint matrix τ^P, which is shown as follows:

$$\min_{G \geq 0} j(\mathbb{R}\colon G, S) = \sum_{R_{o_io_j} \in \mathbb{R}} \|R_{o_io_j} - G_{o_i}S_{o_io_j}G^T_{o_j}\|^2 + \sum_{p=1}^{P} tr\left(G^T \tau^P G\right) \quad . \tag{8}$$

Here, ‖.‖ denotes the Frobenius norm, and *tr* (.) denotes the trace. Our sparse implementation for this 3-*FPNMF* model reduces the missing relational matrix problem with zero values. Our model is more suitable for real-life heterogeneous datasets with missing values, which differs from those of Žitnik and Zupan (2014) in its non-negative sparse implementation. Our proposed 3*FPNMF* – *MKL* model is shown briefly in Figure 1, while a detailed flowchart is represented in Supplementary Figure S1.

2.3 Multiple kernel learning

Next, we introduce the multiple Kernel Learning (MKL) algorithm (Xu et al., 2013) with the hinge loss soft margin, in which the classifier and the kernel combination coefficients are optimized by solving the hinge loss soft margin MKL problem.

After using the 3-*FPNMF* model in the first phase, the approximate sparse relation matrix $\hat{R}_{o_io_j}$ for target object type pairs o_i and o_j is reconstructed as

$$\hat{R}_{o_io_j} = G_{o_i}S_{o_io_j}G^T_{o_j}. \tag{9}$$

Then, to develop kernel fusion, the resulting kernel matrices are generated using the "Kernel Trick": $K(o_i, o_j) = \hat{R}_{o_io_j}.\hat{R}^T_{o_io_j}$. The kernels are further normalized and smoothed using 2-dimensional linear filters.

Given p base-kernels $\mathbb{K} = \{K_1, K_2, \dots, K_p\}$ developed from the reconstructed relational block matrix $\hat{\mathbb{R}} = \{\hat{R}_{o_io_j} | i = 1, \dots, p;\ j = 1, \dots, p\}$, kernel slack variables for the kernel $K_p \in \mathbb{K}$ are defined as the difference between the target margin θ and the SVM dual objective function

$$DSVM\left(K_p, \alpha\right) = \max_{\alpha \in R^N} \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_n \alpha_m y_n y_m K_p\left(x_n, x_m\right)$$

subject to $\sum_{n=1}^{N} y_n \alpha_n = 0, \alpha_n \geq 0, \forall n$. Then, the slack variable is $\zeta_p = \theta - DSVM(K_p, \alpha)$, and the hinge loss is shown as follows:

$$z_p = \ell\left(\zeta_p\right) = \max\left(0, \zeta_p\right). \tag{10}$$

Therefore, the objective function for this hinge loss soft margin MKL algorithm becomes

$$\min_{\theta, \alpha \in Dom(\alpha), \zeta_p} -\theta + \pi \sum_{p=1}^{P} \zeta_p. \tag{11}$$

subject to $DSVM(K_p, \alpha) \geq \theta - \zeta_p$, $\zeta_p \geq 0$, $p = 1, \ldots, P$.

The objective of the aforementioned hinge loss soft margin MKL is to maximize the margin θ while considering the "errors" from the given P-based kernels. The parameter π balances the contribution of the loss term represented by slack variables ζ_p and the margin θ. π should be in the range $\{\pi | \pi \geq 1/P\}$. Otherwise, there is no solution to the proposed problem. Our proposed framework for gene signature detection from heterogeneous data sources using the $3FPNMF - MKL$ model is depicted in Figure 2.

2.4 Determining best combination of class labels using non-matrix factorization and AUC

In biological datasets such as TCGA, the clinical data are made available. This includes patient sample groups, biological subtypes, drug treatment, and survival/prognosis information. In our current study, we obtain accuracies for different combinations of class labels using the non-matrix factorization technique for the case where there were more than two class labels or subtypes. Among them, the combination of class labels, which produces the highest area under curve (AUC), is chosen for the next step (i.e., module detection). Say, q is the specific combination of class labels, which produces the highest AUC. Find $q = \{\exists i, \exists j\} | \{\exists a, \exists b, \exists k\}$ such that

$$AUC_q = arg\,max\left(\forall_{i,j} AUC_{cl_i cl'_j}, \forall_{a,b,k} AUC_{cl_a cl'_{bk}}\right), \tag{12}$$

where cl denotes the left part of the group combination, cl' signifies the right part of any sample group combination, and $i \in \{1, 2, \ldots, (m-1)\}$, $j \in \{(i+1), (i+2), \ldots, m\}$, $a \in \{1, 2, \ldots, m\}$, $b \in \{1, 2, \ldots, m\}$ & $b \neq a$, $k \in \{, 2, \ldots, m\}$, and $k \neq a$ and $k \neq b$.

2.5 Feature clustering and module detection

After selecting the right class-label combination, we extracted the sub-gene expression data consisting of only the selected class labels and then used them for gene module detection and signature identification.

In our procedure, we first evaluated the power of the soft thresholding, which was then applied to evaluate the adjacency matrix using Pearson's correlation. The topological overlap matrix (TOM) similarity score (Ravasz et al., 2002) was computed from the employed adjacency matrix. The TOM score between two nodes (say, i and j) symbolized as $TOM(i, j)$ is defined as follows:

$$TOM(i,j) = \begin{cases} \dfrac{\sum_{v \neq i,j} X(i,v) X(j,v) + X(i,j)}{min\left\{\sum_{v \neq i} X(i,v), \sum_{v \neq j} X(j,v)\right\} - X(i,j) + 1}, & \text{if} i \neq j, \\ 1, & \text{if i = j,} \end{cases} \tag{13}$$

where X denotes the corresponding adjacency matrix containing Boolean entries. The entry of 1 indicates that the corresponding two nodes share the same connection (i.e., direct connection), while the entry of 0 signifies that no direct connection exists between them.

After obtaining the TOM score, we computed the distance/dissimilarity value between genes (i and j) denoted by $dissTOM(i, j)$,which is shown as follows: $dissTOM(i, j) = 1 - TOM(i, j)$. We conducted average linkage clustering on the multi-omics dissimilarity matrix *dissTOM via* considering all potential pairs of genes/features. Finally, the dynamic tree cut technique (Langfelder et al., 2008) was applied on the clustering dendrogram to determine the gene modules. In order to evaluate the quality of the aforementioned clustering, we calculated different cluster validity index measures, *viz.*, cluster coefficient, heterogeneity, Dunn Index, maximum adjacency ratio, centralization, silhouette width, and scaled connectivity.

2.6 Expression signature detection and classifier models

After finding the gene modules, we estimated Pearson's correlation coefficient (PCC) between each gene pair within the resulted modules. For each module, the mean of the correlations for each gene pair within that particular module was obtained. The module with the maximum mean correlation coefficient was elected as a gene signature. Notably, genes with the elected gene signature are differentially expressed between case and control samples. In order to validate the classification performance of the employed gene signature, we utilized the Prediction Analysis of Microarrays (PAM) classifier with 10-fold cross-validation (CV) on the expression sub-data to classify the underlying class labels. The entire procedure was then repeated ten times. Moreover, we calculated the average scores of several classification performance metrics such as sensitivity, specificity, precision, accuracy, and AUC, individually.

2.7 Functional annotation analysis

We carried out KEGG pathway and Gene Ontology (GO) analyses using the Enrichr database (Chen et al., 2013). Notably, GO terms can be categorized into three kinds, *viz.*, biological process (BP), cellular component (CC), and molecular function (MF). Those significant pathways/GO terms with an adjusted p-value less than 0.05 were identified. Meanwhile, literature research studies were also performed to identify disease-related pathways/GO terms.

TABLE 1 Predictive performance of classification for each pairwise class using the proposed method in LAML multi-omics data, where classes 1, 2, and 3 denote "favorable," "intermediate/normal," and "poor," respectively.

	Sensitivity	Specificity	Precision (PPV)	Negative predictive value	Accuracy	AUC
Class 1 vs. Class 2	0.5161	0.6907	0.3478	0.8171	0.6484	0.6202
Class 1 vs. Class 3	0.5484	0.8235	0.7391	0.6667	0.6923	0.7713
Class 1 vs. classes 2 and 3	0.5385	0.3871	0.7865	0.1667	0.5093	0.4608
Class 2 vs. Class 3	0.6289	0.5	0.7821	0.3208	0.5954	0.5215
Class 2 vs. classes 1 and 3	0.5	0.5052	0.4	0.6049	0.5031	0.4863
Class 3 vs. classes 1 and 2	0.5547	0.4848	0.8068	0.2192	0.5404	0.5528
Max	0.6289	0.8235	0.8068	0.8171	0.6923	0.7713

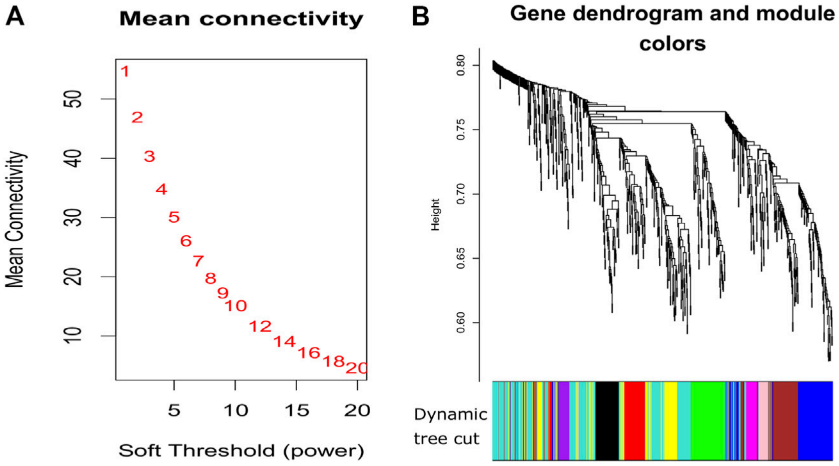


FIGURE 3
Plots for soft thresholding and dendrogram for our proposed method. **(A)** Power computing for soft thresholding and **(B)** dendrogram plots with dynamic tree cut.

3 Results

3.1 Data sources

For our experiment, TCGA acute myeloid leukemia (LAML) multi-omics dataset (https://xenabrowser.net/datapages/?cohort=GDC%20TCGA%20Acute%20Myeloid%20Leukemia%20(LAML)&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443) contained six heterogeneous profiles such as the gene expression (IlluminaGA) profile, DNA methylation (Illumina Methylation 27k) profile, exon expression (IlluminaGA) profile, miRNA profile, pathway activity (Paradigm IPLs) profile, and copy number (GISTIC2) profile. Initially, the gene expression profile included 179 samples and 20,113 genes. For the methylation profile, there are 194 samples and 27,578 methylation probes. Particularly, for the methylation profile, many genes are profiled with more than one probe. In the exon expression profile, there are a total of 219,296 chromosome ids and 179 samples. Here, many genes are connected with more than one chromosome id. The miRNA profile contains 705 miRNAs and 188 samples. The pathway activity profile has 7,203 genes and 173 samples, while the copy number profile consists of 24,776 genes and 191 samples. There are three categories of samples (i.e., class labels) for the LAML multi-omics dataset: i) favorable, ii) intermediate (also called normal), and iii) poor. Specifically, every profile consists of 161 commonly shared LAML samples. Among them, 31 samples belong to the first category, 96 samples are in the second category, and the rest of the samples (= 34) are in the third category. In addition, there are 1,501 uniquely matched genes among the five profiles [i.e., gene expression, DNA methylation, exon expression, pathway activity, and copy number variation (GISTIC2) profiles].

3.2 Statistical validation

First, we selected the sub-data, which contain commonly shared samples (i.e., 161) and genes (i.e., 1,501) for each of the five profiles (i.e., gene expression, DNA methylation, exon expression, pathway activity, and copy number variation profiles). Many matched genes are connected with more than one probe (or chromosome id) for each

TABLE 2 Cluster Validity Index measures of our experiment.

Cluster Validity Index	Score
Dunn Index	0.6461
Average scaled connectivity	0.6834
Silhouette width	−0.0012
Average cluster coefficient	0.2390
Average maximum adjacency ratio	0.2386
Density	0.2327
Centralization	0.1081
Heterogeneity	0.1143

profile. In the case of the miRNA profile, we started to work with the matched samples (n = 161) and all of its miRNAs (n = 705). The empirical Bayes test is performed by limma software on each gene probe or chromosome id for each of the five profiles (i.e., gene expression, DNA methylation, exon expression, pathway activity, and copy number variation profiles) across all the three classes (viz., favorable, intermediate, and poor).

Notably, since there are three classes/groups of samples, here, limma is initially performed between each group pair (i.e., i) favorable vs. intermediate, ii) intermediate vs. poor, and finally iii) favorable vs. poor), then an F-statistics is computed, and finally, the respective p-value is generated from the F-statistics. After the test, for every gene, we only selected the probe or chromosome id with the lowest p-value achieved among all the probes or chromosome ids connected with that gene. As a result, we obtained 728, 272, 1,100, 265, and 904 significant genes for the gene expression, methylation, exon expression, pathway activity, and copy number profiles, respectively. Thereafter, we took the combination of all the significant gene sets, which led to a molecular set of a total of 1,388 genes. Furthermore, the same significance test was applied on each miRNA of the miRNA profile across all the three classes (*viz.*, favorable, intermediate, and poor) as well. We obtained a total of 229 significant miRNAs.

3.3 Expression signature detection and classification

Using the non-matrix factorization technique, we obtained accuracies for different combinations of class labels such as i) Class 1 (favorite) vs. Class 2 (intermediate), ii) Class 1 vs. Class 3 (Poor), iii) Class 1 vs. classes 2 and 3, iv) Class 2 vs. Class 3, v) Class 2 vs. classes 1 and 3, and vi) Class 3 vs. Classes 1 and 2 (as depicted in Table 1). Among them, the second combination, i.e., Class 1 vs. Class 3 produced the highest area under curve (AUC = 0.7713). Hence, we selected the combination for gene signature discovery since other combinations did not produce better AUC scores. After obtaining right combinations of class labels, we first evaluated the power (=1) for soft thresholding (illustrated in Figure 3A), which was then applied to estimate the adjacency matrix through Pearson's correlation score. Then, the TOM score and distance matrix were computed. To determine gene modules, we applied average linkage clustering and dynamic tree cut methodologies. As a result, we generated a total of 10 gene modules. The numbers of participating differentially expressed genes (DEGs) for these 10 gene modules (represented by black, blue, brown, green, magenta, pink, purple, red, turquoise, and yellow colors) were 50, 99, 90, 74, 23, 25, 22, 51, 214, and 80, respectively. The dendrogram is represented in Figure 3B. The corresponding cluster validity indices in that module detection are illustrated in Table 2. The Average silhouette width plot generated during clustering is represented in Supplementary Figure S2. PCC was calculated between each gene pair within each module. The mean correlation scores of the 10 modules (depicted by blue, green, turquoise, magenta, brown, red, yellow, black, purple, and pink colors) were 0.0268, 0.2562, 0.0321, 0.3914, 0.1143, 0.0215, 0.0570, 0.4029, 0.3455, and 0.1605, respectively. The black module had the highest mean correlation coefficient score (= 0.4029 in Table 3). Thus, it was selected as the gene signature. The resultant gene signature contained 50 DEGs (see Table 3). To verify the classification performance of the resultant signature, we applied the PAM classifier through the 10-fold cross-validation (CV) on all the features and samples of signature data in order to classify the groups (favorite and poor). The entire procedure was then repeated 10 times. In the experiment, the mean sensitivity, mean specificity, mean precision, mean accuracy, and mean AUC were 69.12%, 84.19%, 82.79%, 76.31, and 0.8273, respectively (see Figure 4; Table 4). Based on the gene set enrichment analysis on the 50 genes of the signature using the Enrichr web database, we extracted significant KEGG pathway and Gene Ontology (GO) terms. Among the KEGG pathways, the Rap1 signaling pathway (hsa04015) is the most significant pathway (adjusted p-value = 7.497×10^{-06}) that contains eight genes (viz., *EFNA1*, *GNAO1*, *TIAM1*, *CSF1*, *ITGB3*, *ITGA2B*, *THBS1*, and *MAPK13*). Second, the most significant pathway is the PI3K-Akt signaling pathway (hsa04151) with an adjusted

TABLE 3 Feature (gene) names and average (avg.) Pearson's correlation coefficient (PCC) for the pairwise manner within the TCGA LAML signature.

Measure	Value/description
# Features	50
Gene symbols	*HK2, CHRDL1, EFNA1, ARNTL, EIF4A1, MS4A2, BMP2, FHL2, SH2D2A, CSF1, KLRG1, ITGB3, SH3BP5, CCL4, RORA, CAMK2D, BIRC3, TP53, S1PR5, GNAZ, EPOR, TBX21, GATA3, TIAM1, IL2RB, LRIG1, GRAP2, PLEKHA1, THBS1, MAF, IL18RAP, EDN1, ETS1, GATA1, ITGA2B, A2M, LCK, MAPK13, GZMB, PTGDR, MYBL1, RASGRP1, ARG1, PKLR, GNAO1, PRF1, CD8A, FASLG, ABCG2,* and *CCL5*
Average PCC	0.403

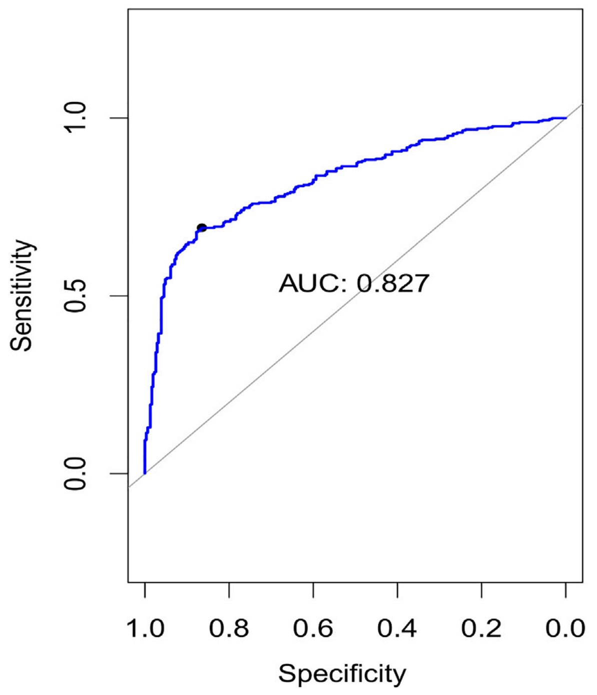


FIGURE 4
Plots of the area under curve (AUC) for 10-fold cross-validation.

TABLE 4 Classification metrics for our experiment.

Evaluation metric	Average score (std)
Precision	0.8279 (±0.027)
Sensitivity	0.6912 (±0.025)
Specificity	0.8419 (±0.028)
Accuracy	0.7631 (±0.0208)
AUC	0.8273

p-value of 1.128 × 10^{-05}, which consists of nine genes (viz., *EFNA1, CSF1, ITGB3, ITGA2B, IL2RB, FASLG, TP53, THBS1*, and *EPOR*). The following eight pathways are the cytokine–cytokine receptor interaction (hsa04060) (adj. p-value = 1.437 × 10^{-05}), inflammatory bowel disease (IBD) (hsa05321) (adj. p-value = 2.1E-05), proteoglycans in cancer (hsa05205) (adj. p-value = 2.1 × 10^{-05}), hematopoietic cell lineage (hsa04640) (adj. p-value = 6.752 × 10^{-05}), T-cell receptor signaling pathway (hsa04660) (adj. p-value = 1 × 10^{-4}), TNF signaling pathway (hsa04668) (adj. p-value = 2 × 10^{-4}), osteoclast differentiation (hsa04380) (adj. p-value = 3 × 10^{-4}), and Ras signaling pathway (hsa04014) (adj. p-value = 3 × 10^{-4}) (also see Table 5). Among the significant GO:BP terms, the positive regulation of cellular metabolic processes (GO:0031325) (adjusted p-value = 8.02947 × 10^{-05}) was ranked as the most significant, which contains six genes (*EDN1*, *CSF1*, *CCL5*, *GATA3*, *THBS1*, and *TP53*). The second most significant GO term is the regulation of inflammatory responses (GO:0050727) with an adjusted p-value of 8.029 × 10^{-05}. This term consists of seven genes (*CCL5, CCL4, RORA, GATA3, ETS1, BIRC3*, and *MAPK13*) (Table 5). Among the significant GO:CC terms, the platelet alpha granule (GO:0031091) (adjusted p-value = 4 × 10^{-3}) contains four genes (viz., *ITGB3, ITGA2B, A2M*, and *THBS1*), while among the GO:MF terms, the core promoter binding factor (GO:0001047) (adjusted p-value = 8 × 10^{-4}) contains five genes (*viz.*, *RORA, GATA3, GATA1, TP53*, and *ARNTL*). For details of the top significant pathways and GO terms, see Table 5.

4 Discussion

Multi-view data integration and gene signature detection are currently the most challenging tasks for biomedical researchers. As different datasets contain different characteristics, integration of data from multi-platforms with significant feature reduction and gene module detection will give a more comprehensive view of how biology unravels at a granular level. Therefore, we introduced the novel approach of multi-platform data integration technique, 3PNMF-MKL, for multi-platform data integration and gene signature detection. This approach applies the integrated utilization of statistical methods, data fusion through three-factor penalized non-negative matrix factorization, and soft margin hinge loss-based multiple kernel learning. We then tested our approach using TCGA LAML multi-omics dataset, which contains five different profiles (viz., gene expression, DNA methylation, exon expression, pathway activity, and copy number). Overall, our algorithm provides excellent AUC (= 0.827) for classifying the class labels for the underlying features (genes) within the chosen gene signature. Furthermore, we performed a functional analysis using the KEGG pathway and Gene Ontology database to interpret those identified relevant feature genes. Collectively, our novel approach is applicable to any kind of multi-modal datasets.

Our proposed method 3PNMF-MKL includes data integration employed by means of differential expression/methylation analysis using limma, non-negative matrix factorization, and soft margin hinge loss, as well as gene signature detection together. 3PNMF-MKL employs the application of best gene module discovery with the help of dynamic linkage clustering, dynamic tree cut, and correlation analysis to achieve the use of best gene module discovery (in terms of gene signature discovery) . So far, there are many state-of-the-art methods available regarding data integration (Yang and Michailidis, 2016; Ray et al., 2017) and gene signature discovery (Cun and Frohlich, 2012; (Zhang and Xiao, 2020), but very few existing methods are recently available where data integration and gene signature detection work together in the same framework (Fujita et al., 2018). We, here, compared our proposed method 3PNMF-MKL with the existing method (Zhang and Xiao, 2020) used for TCGA acute myeloid leukemia dataset. In our proposed method, we obtained a 50-gene signature generated after analyzing multi-omics data integration where the other method (Zhang and Xiao, 2020) produced an eight-gene signature from analyzing the only gene expression data not by multi-omics data integration. Also, we obtained

TABLE 5 Top five significant KEGG pathways and Gene Ontology (GO) terms* for the gene set belonging to the LAML signature.

KEGG pathway name	Gene symbol	Z-score	Adjusted *p*-value
Rap1 signaling pathway (hsa04015)	*EFNA1*, *GNAO1*, *TIAM1*, *CSF1*, *ITGB3*, *ITGA2B*, *THBS1*, and *MAPK13*	−1.961	7.497 × 10−06
PI3K-Akt signaling pathway (hsa04151)	*EFNA1*, *CSF1*, *ITGB3*, *ITGA2B*, *IL2RB*, *FASLG*, *TP53*, *THBS1*, and *EPOR*	−2.041	1.128 × 10−05
Cytokine–cytokine receptor interaction (hsa04060)	*BMP2*, *IL18RAP*, *CSF1*, *CCL5*, *IL2RB*, *CCL4*, *FASLG*, and *EPOR*	−1.829	1.437 × 10−05
Inflammatory bowel disease (IBD) (hsa05321)	*MAF*, *IL18RAP*, *TBX21*, *RORA*, and *GATA3*	−1.858	2.1 × 10−05
Proteoglycans in cancer (hsa05205)	*TIAM1*, *CAMK2D*, *ITGB3*, *FASLG*, *TP53*, *THBS1*, and *MAPK13*	−1.910	2.1 × 10−05
Positive regulation of the cellular metabolic process (GO:BP: GO:0031325)	*EDN1*, *CSF1*, *CCL5*, *GATA3*, *THBS1*, and *TP53*	−1.551	8.029 × 10−05
Regulation of inflammatory response (GO:BP: GO:0050727)	*CCL5*, *CCL4*, *RORA*, *GATA3*, *ETS1*, *BIRC3*, and *MAPK13*	−1.029	8.029 × 10−05
Positive regulation of gene expression (GO:BP: GO:0010628)	*BMP2*, *CSF1*, *TBX21*, *FHL2*, *RORA*, *GATA3*, *ETS1*, *GATA1*, *MYBL1*, *THBS1*, *TP53*, and *ARNTL*	−1.668	8.029 × 10−05
Cytokine-mediated signaling pathway (GO:BP: GO:0019221)	*CAMK2D*, *IL18RAP*, *CSF1*, *CCL5*, *CCL4*, *IL2RB*, *FASLG*, *RORA*, *GATA3*, *TP53*, and *BIRC3*	−1.343	8.029 × 10−05
Positive regulation of nucleic acid-templated transcription (GO:BP: GO: 1903508)	*BMP2*, *TBX21*, *FHL2*, *RORA*, *GATA3*, *ETS1*, *GATA1*, *MYBL1*, *TP53*, and *ARNTL*	−2.001	8.029 × 10−05
Platelet alpha-granule (GO-CC: GO:0031091)	*ITGB3*, *ITGA2B*, *A2M*, and *THBS1*	−1.639	4 × 10−3
Platelet alpha-granule membrane (GO-CC: GO:0031092)	*ITGB3* and *ITGA2B*	−2.148	0.023
Core promoter binding (GO-MF: GO:0001047)	*RORA*, *GATA3*, *GATA1*, *TP53*, and *ARNTL*	−1.279	8 × 10−4
Core promoter sequence-specific DNA binding (GO-MF: GO:0001046)	*RORA*, *GATA3*, *GATA1*, and *TP53*	−1.295	1.9 × 10−3
Transcription regulatory region DNA binding (GO-MF: GO:0044212)	*TBX21*, *RORA*, *GATA3*, *GATA1*, *MYBL1*, *TP53*, and *ARNTL*	−1.322	1.9 × 10−3
Cytokine activity (GO-MF: GO:0005125)	*BMP2*, *EDN1*, *CSF1*, *CCL5*, and *CCL4*	−1.224	2 × 10−3
Transcription factor activity and RNA polymerase II core promoter proximal region sequence-specific binding (GO-MF: GO:0000982)	*GATA3*, *ETS1*, *GATA1*, *MYBL1*, *TP53*, and *ARNTL*	−1.604	2.2 × 10−3

*Gene Ontology (GO) has three domains: biological process (BP), cellular component (CC), and molecular function (MF).

0.87 as the training set's 1-year AUC and 0.72 as the test set's 1-year AUC in the signature survival study (by cox regression), while the other method obtained 0.86 as the training set's 1-year AUC and 0.69 as the test set's 1-year AUC for the gene expression data. Therefore, in all perspectives, our signatures are stronger than the other.

5 Conclusion and future directions

No method, which deals with data integration non-matrix factorization, soft margin hinge loss, and gene signature together, exists in the field of bioinformatics, whereas our work is concerned with the process of integration of multi-omics data employing multi-dimensional schemes such as differential expression/methylation analysis using limma, non-negative matrix factorization, soft margin hinge loss, and gene signature detection through the use of best gene module discovery using dynamic linkage clustering, dynamic tree cut method, and correlation analysis, respectively. The achievement of a high classification accuracy of 0.8273 also represents superior performance for our proposed algorithm. In addition, our method outperformed the state-of-the-art methods in terms of computing AUC. Expansion of our current approach with a deep learning strategy to tackle the integrative problem at a single-cell level is our future directive. In future work, we will collaborate with a wet laboratory to validate our experimental results in order to make it more promising.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

SM and AS formulated the problem and conceived the design of the study. SM, AS, and SN performed the experimental analysis. SD, SG, SP, UM, and ZZ wrote the manuscript. All authors contributed in editing and revising the manuscript.

Funding

ZZ was partially supported by the Cancer Prevention and Research Institute of Texas (CPRIT RP170668 and RP180734) (to ZZ). Publication costs were funded by ZZ's Professorship Fund. The funder had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1095330/full#supplementary-material

SUPPLEMENTARY FIGURE S1
Detailed flowchart of the proposed 3PNMF-MKL framework.

SUPPLEMENTARY FIGURE S2
Average silhouette width during clustering.

References

Bandyopadhyay, S., and Mallik, S. (2016). Integrating multiple data sources for combinatorial marker discovery: A study in tumorigenesis. *IEEE/ACM Trans. Comput. Biol. Bioinform* 15, 673–687. doi:10.1109/TCBB.2016.2636207

Bandyopadhyay, S., Mallik, S., and Mukhopadhyay, A. (2013). A survey and comparative study of statistical tests for identifying differential expression from microarray data. *IEEE/ACM Trans. Comput. Biol. Bioinform* 11, 95–115. doi:10.1109/TCBB.2013.147

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinforma.* 14, 128. doi:10.1186/1471-2105-14-128

Cun, Y., and Frohlich, H. (2012). Biomarker gene signature discovery integrating network knowledge. *Biol. (Basel)* 1, 5–17. doi:10.3390/biology1010005

Fujita, N., Mizuarai, S., Murakami, K., and Nakai, K. (2018). Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Sci. Rep.* 8, 9743. doi:10.1038/s41598-018-28066-w

Gaur, L., Bhandari, M., Razdan, T., Mallik, S., and Zhao, Z. (2022). Explanation-driven deep learning model for prediction of brain tumour status using mri image data. *Front. Genet.* 448, 822666. doi:10.3389/fgene.2022.822666

Ghose, P., Alavi, M., Tabassum, M., Uddin, M. A., Biswas, M., Mahbub, K., et al. (2022). Detecting Covid-19 infection status from chest x-ray and ct scan via single transfer learning-driven approach. *Front. Genet.* 13, 980338. doi:10.3389/fgene.2022.980338

Henry, V. J., Bandrowski, A. E., Pepin, A.-S., Gonzalez, B. J., and Desfeux, A. (2014). Omictools: An informative directory for multi-omic data analysis. *Database* 2014, bau069. doi:10.1093/database/bau069

Imielinski, M., Cha, S., Rejtar, T., Richardson, E. A., Karger, B. L., and Sgroi, D. C. (2012). Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Mol. Cell. Proteomics* 11, M111.014910. doi:10.1074/mcp.M111.014910

Kandimalla, R., Shimura, T., Mallik, S., Sonohara, F., Tsai, S., Evans, D. B., et al. (2022). Identification of serum mirna signature and establishment of a nomogram for risk stratification in patients with pancreatic ductal adenocarcinoma. *Ann. Surg.* 275, e229–e237. doi:10.1097/SLA.0000000000003945

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for r. *Bioinformatics* 24, 719–720. doi:10.1093/bioinformatics/btm563

Li, P., Guo, M., and Sun, B. (2019). Integration of multi-omics data to mine cancer-related gene modules. *J. Bioinforma. Comput. Biol.* 17, 1950038. doi:10.1142/S0219720019500380

Mallik, S., Bhadra, T., and Maulik, U. (2017). Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data. *IEEE Trans. Nanobioscience* 16, 3–10. doi:10.1109/TNB.2017.2650217

Mallik, S., and Zhao, Z. (2020). Graph-and rule-based learning algorithms: A comprehensive review of their applications for cancer type classification and prognosis using genomic data. *Briefings Bioinforma.* 21, 368–394. doi:10.1093/bib/bby120

Maulik, U., Mallik, S., Mukhopadhyay, A., and Bandyopadhyay, S. (2015). Analyzing large gene expression and methylation data profiles using statbicrm: Statistical biclustering-based rule mining. *PLoS One* 10, e0119448. doi:10.1371/journal.pone.0119448

Mo, Q., Wang, S., Seshan, V. E., Olshen, A. B., Schultz, N., Sander, C., et al. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proc. Natl. Acad. Sci.* 110, 4245–4250. doi:10.1073/pnas.1208949110

Pellet, J., Lefaudeux, D., Royer, P.-J., Koutsokera, A., Bourgoin-Voillard, S., Schmitt, M., et al. (2015). A multi-omics data integration approach to identify a predictive molecular signature of clad. *Eur. Respir. J.* 46, OA3271. doi:10.1183/13993003.congress-2015.OA3271

Qiu, C., Yu, F., Su, K., Zhao, Q., Zhang, L., Xu, C., et al. (2020). Multi-omics data integration for identifying osteoporosis biomarkers and their biological interaction and causal mechanisms. *Iscience* 23, 100847. doi:10.1016/j.isci.2020.100847

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi:10.1126/science.1073374

Ray, B., Liu, W., and Fenyo, D. (2017). Adaptive multiview nonnegative matrix factorization algorithm for integration of multimodal biomedical data. *Cancer Inf.* 16, 1176935117725727. doi:10.1177/1176935117725727

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids Res.* 43, e47. doi:10.1093/nar/gkv007

Saeed, S., Haroon, H. B., Naqvi, M., Jhanjhi, N. Z., Ahmad, M., and Gaur, L. (2022). "A systematic mapping study of low-grade tumor of brain cancer and csf fluid detecting approaches and parameters," in *Approaches and applications of deep learning in virtual medical care*, 236–259. doi:10.4018/978-1-7998-8929-8.ch010

Serra, A., Fratello, M., Fortino, V., Raiconi, G., Tagliaferri, R., and Greco, D. (2015). Mvda: A multi-view genomic data integration methodology. *BMC Bioinforma.* 16, 261. doi:10.1186/s12859-015-0680-3

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 3, 3. doi:10.2202/1544-6115.1027

Wang, D., and Gu, J. (2016). Integrative clustering methods of multi-omics data for molecule-based cancer classifications. *Quant. Biol.* 4, 58–67. doi:10.1007/s40484-016-0063-4

Xu, X., Tsang, I. W., and Xu, D. (2013). Soft margin multiple kernel learning. *IEEE Trans. neural Netw. Learn. Syst.* 24, 749–761. doi:10.1109/TNNLS.2012.2237183

Yang, Z., and Michailidis, G. (2016). A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 32, 1–8. doi:10.1093/bioinformatics/btv544

Zhang, B., and Kuster, B. (2019). Proteomics is not an island: Multi-omics integration is the key to understanding biological systems. *Mol. Cell. Proteomics* 18, S1–S4. doi:10.1074/mcp.E119.001693

Zhang, Y., and Xiao, L. (2020). Identification and validation of a prognostic 8-gene signature for acute myeloid leukemia. *Leukemia Lymphoma* 61, 1981–1988. doi:10.1080/10428194.2020.1742898

Žitnik, M., and Zupan, B. (2014). Data fusion by matrix factorization. *IEEE Trans. pattern analysis Mach. Intell.* 37, 41–53. doi:10.1109/TPAMI.2014.2343973

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 20 April 2023
DOI 10.3389/fgene.2023.1154120

Check for updates

OPEN ACCESS

EDITED BY
Zhibin Lv,
Sichuan University, China

REVIEWED BY
Lei Wang,
Changsha University, China
Yansu Wang,
University of Electronic Science and Technology of China, China

*CORRESPONDENCE
Saurav Mallik,
✉ sauravmtech2@gmail.com,
✉ smallik@hsph.harvard.edu
Hong Qin,
✉ hong-qin@utc.edu

RECEIVED 30 January 2023
ACCEPTED 04 April 2023
PUBLISHED 20 April 2023

CITATION
Rout RK, Umer S, Khandelwal M, Pati S, Mallik S, Balabantaray BK and Qin H (2023), Identification of discriminant features from stationary pattern of nucleotide bases and their application to essential gene classification.
Front. Genet. 14:1154120.
doi: 10.3389/fgene.2023.1154120

Identification of discriminant features from stationary pattern of nucleotide bases and their application to essential gene classification

Ranjeet Kumar Rout[1], Saiyed Umer[2], Monika Khandelwal[1], Smitarani Pati[3], Saurav Mallik[4,5]*, Bunil Kumar Balabantaray[6] and Hong Qin[7]*

[1]National Institute of Technology Srinagar, Hazratbal, Jammu and Kashmir, India, [2]Aliah University, Kolkata, West Bengal, India, [3]Dr. B R Ambedkar National Institute of Technology Jalandhar, Jalandhar, Punjab, India, [4]Harvard T H Chan School of Public Health, Boston, United States, [5]Department of Pharmacology and Toxicology, University of Arizona, Tucson, AZ, United States, [6]National Institute of Technology Meghalaya, Shillong, Meghalaya, India, [7]Department of Computer Science and Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, United States

Introduction: Essential genes are essential for the survival of various species. These genes are a family linked to critical cellular activities for species survival. These genes are coded for proteins that regulate central metabolism, gene translation, deoxyribonucleic acid replication, and fundamental cellular structure and facilitate intracellular and extracellular transport. Essential genes preserve crucial genomics information that may hold the key to a detailed knowledge of life and evolution. Essential gene studies have long been regarded as a vital topic in computational biology due to their relevance. An essential gene is composed of adenine, guanine, cytosine, and thymine and its various combinations.

Methods: This paper presents a novel method of extracting information on the stationary patterns of nucleotides such as adenine, guanine, cytosine, and thymine in each gene. For this purpose, some co-occurrence matrices are derived that provide the statistical distribution of stationary patterns of nucleotides in the genes, which is helpful in establishing the relationship between the nucleotides. For extracting discriminant features from each co-occurrence matrix, energy, entropy, homogeneity, contrast, and dissimilarity features are computed, which are extracted from all co-occurrence matrices and then concatenated to form a feature vector representing each essential gene. Finally, supervised machine learning algorithms are applied for essential gene classification based on the extracted fixed-dimensional feature vectors.

Results: For comparison, some existing state-of-the-art feature representation techniques such as Shannon entropy (SE), Hurst exponent (HE), fractal dimension (FD), and their combinations have been utilized.

Discussion: An extensive experiment has been performed for classifying the essential genes of five species that show the robustness and effectiveness of the proposed methodology.

KEYWORDS

essential genes, DNA, co-occurrence matrix, feature analysis, classification

1 Introduction

Essential genes are necessary for the survival of a living being and are considered the basis of life. Essential genes consist of vital data of genomes and, hence, could be the key to the broad interpretation of life and expansion (Juhas et al., 2011). It decides significant attributes involving cellular structure, chemistry, and reproduction, among others. Genomes have encoded data for the functions regularly viewed as in all life forms, and the instructions could be species-specific. Some genes appear essential for survival, whereas others seem to be optional. Essential genes have been provided to segregate genes and determine the fundamental sustaining cellular life components. Deletion of an essential gene would result in cell death. As a result, essential gene prediction aids in identifying the bare minimum of genes necessary for the vital survival of specific cell types. The discovery and analysis of essential genes aids our understanding of origin of life (Koonin, 2000). Furthermore, essential genes play a crucial role in synthetic molecular biology, vital to genome development. An extensive comprehension of essential genes can empower researchers to clarify the biological essence of microorganisms (Juhas et al., 2014), generate the smallest genome subset (Itaya, 1995), evolve promising medication targets, and create probable drugs to fight infectious diseases (Dickerson et al., 2011). Due to their significance, the identification of essential genes has been viewed as essential in bioinformatics and genomics.

Essential genes are a set of genes necessary for an organism to thrive in a certain climate. Most of these are only necessary for particular circumstances. For instance, if a cell is supplied with the amino acid lysine, the gene responsible for lysine production is non-essential. However, if the amino acid supply is unavailable, the gene encoding the enzyme responsible for lysine biosynthesis becomes essential, as protein synthesis is not possible without it. Essential genes regulate the activity of fundamental cells in almost every species (Qin, 2019; Guo et al., 2021). Genes are essential if they cannot be knocked out individually under circumstances when most of the needed nutrients are present in the growth medium and the organism grows at its optimal temperature. One of the major issues is determining which identified genes are necessary. There are various experimental techniques to identify essential genes in microorganisms, such as gene knockouts (Roemer et al., 2003), RNA interference (Cullen and Arndt, 2005), transposon mutagenesis (Veeranagouda et al., 2014), and single-gene knockout procedures (Giaever et al., 2002). However, these experimental techniques have various benefits and are generally good. They are still expensive and laborious. So, there is a need for computational methods to identify essential genes.

Because essential genes have biological significance, several computational methods, particularly machine learning methods, have been employed to ascertain them. For this objective, many feature extraction and model building approaches have been developed (Gil et al., 2004; McCutcheon and Moran, 2010; Juhas et al., 2012; Mobegi et al., 2017). Chen and Xu (2005) effectively used high-throughput data and machine learning techniques in *Saccharomyces cerevisiae* to evaluate protein dispensability. Seringhaus et al. (2006) constructed a machine learning model to predict essential genes in *S. cerevisiae* using several intrinsic genomic factors. Additionally, Yuan et al. (2012) designed three machine learning techniques based on informative genomic characteristics to detect knockdown lethality in mice. Deng (2015) proposed an important gene classification algorithm using hybrid characteristics like intrinsic and context-dependent genome aspects. This model acquired area under the receiver operating characteristic curve (AUC) scores of 0.86–0.93 when testing the same organism and scores of 0.69–0.89 when predicting cross-organisms using ten-fold cross-validation.

Zhang et al. (2020) have contributed significantly by combining sequence- and network-based features to identify essential genes and arrived at valid results by utilizing a deep learning-based model to learn the characteristics generated from sequencing data and protein–protein interaction networks. Liu et al. (2017) published the findings of comprehensive research on 31 bacterial species, including cross-validation, paired, self-test, and leave-one-species-out experiments. Rout et al. (2020) proposed a method to identify essential genes of four species based on various quantitative methods, including purine and pyrimidine distribution. Le et al. (2020) proposed a model for identifying essential genes using an ensemble deep neural network. Xu et al. (2020) developed a method to predict essential genes in prokaryotes based on sequence-based features using an artificial neural network. A web server, Human Essential Genes Interactive Analysis Platform (HEGIAP), was developed by Chen et al. (2020) for detailed analysis of human essential genes.

An expression-based predictor was developed by Kuang et al. (2021) to recognize the essential genes in humans. The predictor utilized gene expression profiles to predict lncRNAs in cancer cells. Senthamizhan et al. (2021) created a database NetGenes for essential genes, which contains predictions for 2,711 bacterial species using network-based features. The protein–protein interaction network was used to extract features from the STRING database. Marques de Castro et al. (2022)predicted the essential genes in Tribolium castaneum and Drosophila melanogaster based on the physicochemical and statistical data along with subcellular locations. They extracted extrinsic and intrinsic attributes from the essential and nonessential data. This paper analyzed the DNA sequences of five species, i.e., Homo sapiens, Danio rerio, D. melanogaster, Mus musculus, and Arabidopsis thaliana, to identify essential genes. The proposed model extracts co-occurrence matrices from the essential gene sequences to find some informative patterns that distinguish the species. This paper also finds the impact of different co-occurrence matrices and existing features, such as Hurst exponent (HE), fractal dimension (FD), Shannon entropy (SE), and modified Shannon entropy (MSE).

The rest of the paper is structured in the following manner. The definitions of various fundamental parameters are given in Section 2, with relevant descriptions. The proposed methodology with detailed dataset description is discussed in Section 3. The efficiency of our strategy is proven by experimental findings and comments in Section 4, which summarizes the paper by highlighting the most important aspects of the whole investigation. Finally, the paper is concluded in Section 5.

2 Basic terminology

Essential genes are a family linked to critical cellular activities for survival of species. Identifying essential genes is a multidisciplinary process that necessitates both computational and wet-lab validation experiments. Several machine learning methods have been developed to improve classification accuracy, making it a time-consuming and resource-intensive process. Hence, with lower validation costs, most of these methods use supervised methods, which necessitate massive labeled training data sets, typically impractical for less-sequenced species. On the other hand, the rise of high-throughput wet-lab experimental approaches like next-generation sequencing has resulted in an oversupply of unlabeled essential gene sequence data. In the initial study, it has been observed that a fixed-dimensional feature vector represents every DNA sequence by using various quantitative measures, such as SE, MSE, FD, and HE. To estimate these quantitative measures, we convert gene sequences into binary sequences based on pyrimidine and purine distribution. The two main forms of nucleotide bases in DNA are made up of nitrogenous bases. Adenine (A) and guanine (G) are purines, whereas cytosine (C) and thymine (T) are pyrimidines. Here, purine and pyrimidine bases are expressed as 1 and 0, respectively.

$$A/G \rightarrow 1 \ and \ \mathrm{C}/T \rightarrow 0. \tag{1}$$

2.1 Shannon entropy and modified Shannon entropy

SE may be used to determine how much uncertainty or information a sequence contains (Zurek, 1989; Khandelwal et al., 2022b). The uncertainty affects the distribution of each word. A sequence's uncertainty concerning a base pair ranges from 0 to 2n, where n is the length of a word. The SE uses the probability p of the two possibilities (0/1) to calculate information entropy. The following equation gives the SE of a binary sequence:

$$\mathrm{SE} = -\sum_{i=0}^{1} p_i \, log_2(p_i), \tag{2}$$

where p_i indicates the probability of two values regarding the binary sequence, and SE is used to compute the uncertainty in a binary string (Khandelwal et al., 2022a). When the probability p = 0, the event is assured never to happen, resulting in no uncertainty and entropy of 0. Similarly, if p = 1, the result is definite; hence, the entropy must be 0. When p = 1/2, the uncertainty is highest, and the SE is 1. The MSE of different word size is given by

$$MSE = -\sum_{j=1}^{k} w_j \, log_2\left(w_j\right), \tag{3}$$

where w_j indicates the frequency of the j^{th} word in the gene sequence. For instance, for a word of length 1, w_j is determined using the frequencies of purine or pyrimidine 0, 1, and for a word of length 2, w_j is determined using the two-time repeat of purine or pyrimidine 00, 10, 01, and 11. The number of words determined by taking the maximum length of both purines and pyrimidines is represented by k (Rout et al., 2020).

2.2 Hurst exponent

The HE evaluates a data set's smoothness and degree of similarities. The HE is often used to analyze auto-correlation in time-series analysis. It is calculated using rescaled range analysis (R/S analysis) and has a value of 0–1 (Hurst, 1951; Khandelwal et al., 2022c). A negative auto-correlation of a time series is indicated by a HE value between 0 and 0.5, while a HE value between 0.5 and 1 indicates a positive auto-correlation. If the HE value is 0.5, the series is random, meaning that there is no relation between the variable and its previous values (Hassan et al., 2021; Rout et al., 2022). The HE of a binary sequence D_n is computed by the following equation:

$$\frac{R(n)}{S(n)} = \left(\frac{n}{2}\right)^{HE}, \tag{4}$$

where

$$S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i - m)^2}, \tag{5}$$

and

$$R(n) = max(X_1, X_2, \ldots, X_n) - min(X_1, X_2, \ldots, X_n), \tag{6}$$

$$X_t = \sum_{i=1}^{t}(D_i - m) \quad for \ t = 1, 2, 3, \ldots, n \tag{7}$$

$$m = \frac{1}{n}\sum_{i=1}^{n} D_i. \tag{8}$$

2.3 Fractal dimension

Every DNA sequence is converted into indicator matrices (Rout et al., 2018; Umer et al., 2021). Let X = {A, T, C, and G} denote the set of finite alphabet nucleotides, and D(N) denote a DNA sequence with four symbols from X of length N. The indicator function for every DNA sequence is described by the following equation:

$$F\colon D(N) \times D(N) \rightarrow \{0,1\}, \ and \ D(N) = \{0,1\}, \tag{9}$$

such that the indicator matrix will be

$$I(N,N) = \begin{cases} 1, & \text{if } s_i = s_j \\ 0, & \text{if } s_i \neq s_j \end{cases} \quad where \ s_i, s_j \in D(N). \tag{10}$$

Here, I(N, N) is a matrix with values 0 and 1, and it produces a binary image of the DNA sequence as a 2D dot-plot. Within the same sequence, the binary image can represent the distribution of 0s and 1s. It is possible to assign a white dot to 0 and a black dot to 1. The FD from an indicator matrix can be computed as the average number of $\sigma(n)$ of 1, randomly selected n× n from an N× N indicator matrix (Cattani, 2010; Rout et al., 2014; Upadhayay et al., 2019). Using $\sigma(n)$, the FD is computed by the following equation:

$$FD = -\frac{1}{N}\sum_{n=2}^{N}\frac{log(\sigma(n))}{logn}. \tag{11}$$

3 Proposed scheme

In this paper, we used the Database of Essential Genes (http://www.essentialgene.org/) for experimental findings and discussion. This dataset consists of essential genes of five species. There are

TABLE 1 List of species considered in the proposed technique.

Name	Symbol used
Arabidopsis thaliana	AT
Drosophila melanogaster	DOM
Danio rerio	DR
Homo sapiens	HS
Mus musculus	MM
Naming convention for *Arabidopsis thaliana*	$[AT_1 - AT_{356}]$
Naming convention for *Drosophila melanogaster*	$[DOM_1 - DOM_{339}]$
Naming convention for *Danio rerio*	$[DR_1 - DR_{315}]$
Naming convention for *Homo sapiens*	$[HS_1 - HS\ _{2051}]$
Naming convention for *Mus musculus*	$[MM_1 - MM_{125}]$

TABLE 2 Possible sets of occurrences of nucleobases *A, C, T, G* in a DNA sequence or essential gene formed by the combination of vectors, where *I, J, K, L, M, N, O, P* are the co-occurrence matrices.

X	Y	$X^T \times Y$
$X_1 = (A, C, T, G)$	(A, C, T, G)	$\underset{4\times4}{I} = \underset{4\times1}{X_1^T} \times \underset{1\times4}{Y}$
$X_2 = (AA, CC, TT, GG)$	(A, C, T, G)	$\underset{4\times4}{J} = \underset{4\times1}{X_2^T} \times \underset{1\times4}{Y}$
$X_3 = (AC, AT, AG, CT, CG, TG)$	(A, C, T, G)	$\underset{6\times4}{K} = \underset{6\times1}{X_3^T} \times \underset{1\times4}{Y}$
$X_4 = (CA, TA, GA, TC, GC, GT)$	(A, C, T, G)	$\underset{6\times4}{L} = \underset{4\times1}{X_4^T} \times \underset{1\times4}{Y}$
$X_5 = (ACT, ACG, ATG, CTG)$	(A, C, T, G)	$\underset{4\times4}{M} = \underset{4\times1}{X_5^T} \times \underset{1\times4}{Y}$
$X_6 = (CAT, CAG, TAG, TCG)$	(A, C, T, G)	$\underset{4\times4}{N} = \underset{4\times1}{X_6^T} \times \underset{1\times4}{Y}$
$X_7 = (ATC, AGC, AGT, CGT)$	(A, C, T, G)	$\underset{4\times4}{O} = \underset{4\times1}{X_7^T} \times \underset{1\times4}{Y}$
$X_8 = (TCA, GCA, GTA, GTC)$	(A, C, T, G)	$\underset{4\times4}{P} = \underset{4\times1}{X_8^T} \times \underset{1\times4}{Y}$

2,051 *H. sapiens* (HS), 315 *D. rerio* (DR), 339 *D. melanogaster* (DOM), 356 *A. thaliana* (AT), and 125 *M. musculus* (MM) essential genes. Table 1 lists some of the terminologies employed in the proposed technique for reference.

3.1 Proposed feature representation technique

The DNA (deoxyribonucleic acid) sequence of essential genes $\mathcal{S}$ is composed of four bases: adenine (A), guanine (G), cytosine (C), and thymine (T). So, several occurrences may exist with combinations of *A, C, T, G* within the sequence $\mathcal{S}$. The co-occurrences of *A, C, T, G* in the DNA sequence establishes the relationship between the nucleotide. It is the first time that a method has been proposed for finding the co-occurrences of nucleotides *A, C, T, G* within $\mathcal{S}$. The objective of finding these co-occurrences is to analyze the patterns of *A, C, T, G* within the DNA sequence $\mathcal{S}$ to derive some useful features that uniquely discriminate the species by the feature representation of their essential genes. Assuming $x = (A, C, T, G)$ is a vector of the nucleotides, then the possibility of arrangement of these characters in the DNA gene sequences is represented through co-occurrence matrices formed by the vector combination, which are shown in Table 2.

TABLE 3 Co-occurrence matrix *I* that contains several patterns of *A, C, T, G* nucleobases in DNA gene sequence $\mathcal{S}$

	A	C	T	G
A	#(AA)	#(AC)	#(AT)	#(AG)
C	#(CA)	#(CC)	#(CT)	#(CG)
T	#(TA)	#(TC)	#(TT)	#(TG)
G	#(GA)	#(GC)	#(GT)	#(GG)

TABLE 4 Features extracted from a co-occurrence matrix $\mathcal{G}$ of DNA sequence *S*.

Feature	Formulae
Energy	$\sum_{r=0}^{q}\sum_{s=0}^{q}\mathcal{G}'(r,s)^2$
Entropy	$\sum_{r=0}^{q}\sum_{s=0}^{q} - \mathcal{G}'(r,s) \times \ln(\mathcal{G}'(r,s))$
Homogeneity	$\sum_{r=0}^{q}\sum_{s=0}^{q}\frac{\mathcal{G}'(r,s)}{(1+(r-s)^2)}$
Contrast	$\sum_{r=0}^{q}\sum_{s=0}^{q}\mathcal{G}'(r,s) \times (r-s)^2$
Dissimilarity	$\sum_{r=0}^{q}\sum_{s=0}^{q}\mathcal{G}'(r,s) \times \vert(r-s)\vert$

Here, the computed co-occurrence matrices of different combinations of nucleobases represent the distribution of nucleobases throughout the essential gene S. This distribution of nucleobases examines the texture pattern and considered the spatial relationship of nucleobases in the essential gene S. Experimentally, it has been observed that the occurrences of the spatial relationship of nucleobases cannot provide fixed information of the stationary and non-stationary patterns of A, C, T, and G. However, the obtained spatial relationship contains the information of both these patterns at a time. Hence, statistically it is easier to compute information considering both stationary and non-stationary patterns at a time rather than differentiating stationary and non-stationary patterns in S. The essential genes are very critical for the survival of any organism. It is beneficial for cell growth. Each gene sequence is variable in length, and the arrangements *A, C, T, G* nucleobases are zigzag. Hence, finding the stationary and non-stationary patterns of *A, C, T, G* and the co-occurrences of the different combinations of these nucleobases will help find its natural pattern in the gene. Hence, deriving the valuable patterns of the variety of *A, C, T, G* through co-occurrence matrix descriptors will considerably improve the retrieval performance and be eligible to analyze the statistical and structural information effectively from those patterns. Hence, inspired by the co-occurrence matrix of texture analysis (Umer et al., 2016) of image processing and pattern recognition, we have employed the ideas of gray-level co-occurrence matrix. Here, we have computed several co-occurrence matrices from each essential gene data. Now, $\underset{4\times4}{I}$, $\underset{4\times4}{J}$, $\underset{6\times4}{K}$, $\underset{6\times4}{L}$, $\underset{4\times4}{M}$, $\underset{4\times4}{N}$, $\underset{4\times4}{O}$, and $\underset{4\times4}{P}$ co-occurrences

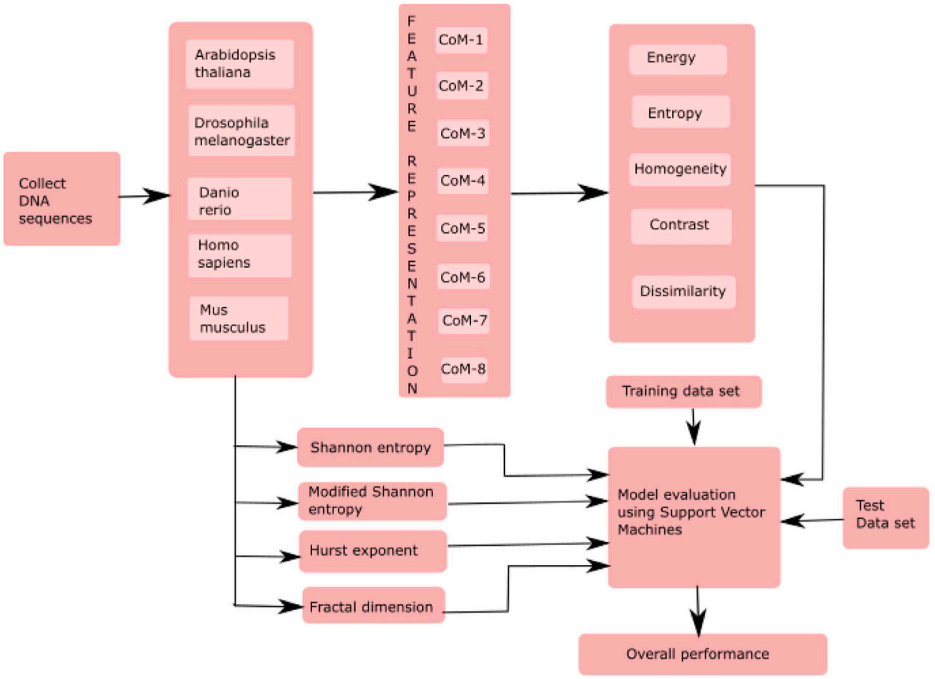


FIGURE 1
Framework of the proposed model for the classification of essential genes. Here, CoM indicates the co-occurrence matrices.

TABLE 5 Demonstration of actual files containing gene sequences corresponding to AT, DOM, DR, HS, and MM species.

	Actual files	Actual files containing DNA sequences
AT	356	356
DOM	339	339
DR	315	315
HS	2054	2051
MM	411	125

matrices are computed that contain several patterns of A, C, T, G nucleobases in each DNA sequence $\mathcal{S}$. These co-occurrence matrices are defined in Table 3, Supplementary Table S1, Supplementary Table S2, Supplementary Table S3, Supplementary Table S4, Supplementary Table S5, Supplementary Table S6, and Supplementary Table S7, respectively.

Here, from the given DNA sequence $\mathcal{S}$, the aforementioned co-occurrence matrices are obtained. Each co-occurrence matrix $\mathcal{G}$ contains the number of occurrences of A, C, T, G nucleobases with a specific combinations and offset in $\mathcal{S}$. Since a sequence $\mathcal{S}$ with q different combinations of A, C, T, G nucleobases will produce a co-occurrence matrix of size $q \times 4$ for the given offset, so the $(r,s)^{th}$ value of a co-occurrence matrix (Table 3, Supplementary Table S1, Supplementary Table S2, Supplementary Table S3, Supplementary Table S4, Supplementary Table S5, Supplementary Table S6, and Supplementary Table S7) gives the number of times that r^{th} and s^{th} nucleobases present in $\mathcal{S}$. Hence, mathematically, here each co-occurrence matrix (Table 3, Supplementary Table S1, Supplementary Table S2, Supplementary Table S3, Supplementary Table S4, Supplementary Table S5, Supplementary Table S6, and Supplementary Table S7) is given by

$$\mathcal{G} = \sum_{i=1}^{n}\sum_{j=1}^{n}\begin{cases}1 & G_{(i,j)} = r \;\&\; G_{(i+\triangle i, j+\triangle j)} = s\\ 0 & \text{otherwise}\end{cases}, \qquad (12)$$

The offset $(\triangle i, \triangle j)$ defines the spatial relation for which the matrix $\mathcal{G}$ is calculated. The number of co-occurrences of the combinations of A, C, T, G present in $\mathcal{S}$ is obtained by the co-occurrence matrices. So, to extract distinguish and discriminant features, each matrix $\mathcal{G}$ is normalized to $\mathcal{G}' = \frac{\mathcal{G}}{\sum_{r=0}^{q}\sum_{s=0}^{q}\mathcal{G}(r,s)}$. Then, the normalized co-occurrence matrix $\mathcal{G}'$ is used to compute some features like entropy, dissimilarity, energy, homogeneity, and contrast. The mathematical definitions of these features are shown in Table 4.

Now, the features defined in Table 4 are extracted from each co-occurrence matrix (Table 3, Supplementary Table S1, Supplementary Table S2, Supplementary Table S3, Supplementary Table S4, Supplementary Table S5, Supplementary Table S6, and Supplementary Table S7), and the list of feature vectors extracted from these matrices is obtained as follows:

$f_I = (f_1, f_2, f_3, f_4, f_5)$ from I (Table 3)
$f_J = (f_6, f_7, f_8, f_9, f_{10})$ from J (Supplementary Table S1)
$f_K = (f_{11}, f_{12}, f_{13}, f_{14}, f_{15})$ from K (Supplementary Table S2)
$f_L = (f_{16}, f_{17}, f_{18}, f_{19}, f_{20})$ from L (Supplementary Table S3)
$f_M = (f_{21}, f_{22}, f_{23}, f_{24}, f_{25})$ from M (Supplementary Table S4)
$f_N = (f_{26}, f_{27}, f_{28}, f_{29}, f_{30})$ from N (Supplementary Table S5)
$f_O = (f_{31}, f_{32}, f_{33}, f_{34}, f_{35})$ from O (Supplementary Table S6)
$f_P = (f_{36}, f_{37}, f_{38}, f_{39}, f_{40})$ from P (Supplementary Table S7)

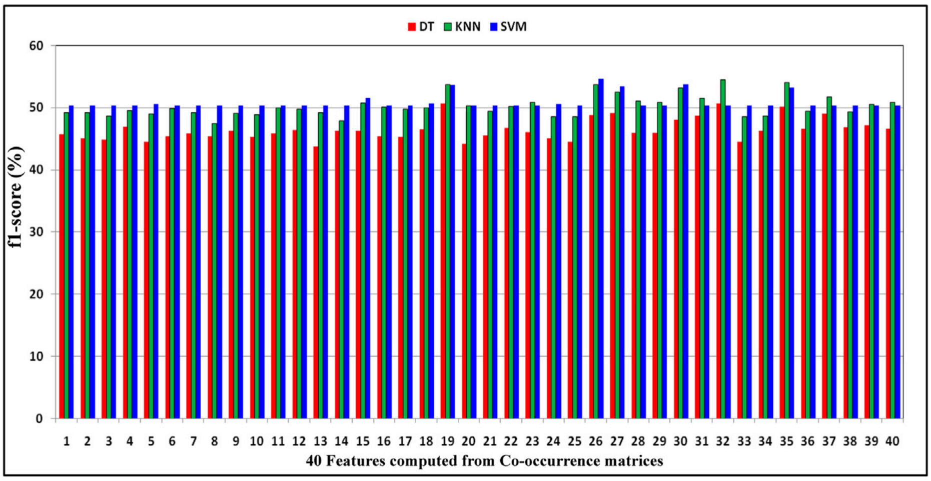


FIGURE 2
Demonstration of distribution of F1-score performance obtained by decision tree, KNN, and SVM classifiers with respect to the 40 features computed from co-occurrence matrices of DNA gene sequence *S*.

TABLE 6 Impact of different co-occurrence features on the classification of essential gene sequences of AT, DOM, DR, HS, and MM species.

Classifier	Accuracy	Precision	Recall	F1-score
Effect of entropy features				
K-nearest neighbors	63.56	56.68	63.56	59.39
Decision tree	52.95	53.56	52.95	53.25
Support vector machine	64.37	41.44	64.37	50.42
Effect of dissimilarity features				
K-nearest neighbors	62.96	57.38	62.96	59.55
Decision tree	52.70	53.84	52.70	53.25
Support vector machine	67.07	58.80	67.07	56.75
Effect of energy features				
K-nearest neighbors	59.48	52.71	59.48	55.46
Decision tree	48.65	49.82	48.65	49.22
Support vector machine	64.94	50.32	64.94	51.83
Effect of homogeneity features				
K-nearest neighbors	63.06	57.59	63.06	59.99
Decision tree	53.61	54.81	53.61	54.19
Support vector machine	67.67	60.76	67.67	58.29
Effect of contrast features				
K-nearest neighbors	64.25	58.92	64.25	61.02
Decision tree	54.80	56.27	54.80	55.51
Support vector machine	68.36	59.82	68.36	58.85

Hence, the final feature representation of a DNA sequence or essential gene $\mathcal{S}$ is given by the feature vector $f = (f_I, f_J, f_K, f_L, f_M, f_N, f_O, f_P)$.

3.2 Classification

In this study, for the classification of the essential genes in the employed species, the decision tree (DT), k-nearest neighbor (KNN), and support vector machine (SVM) classifiers are used. During experimentation, the datasets of each species *Arabidopsis thaliana* (AT), *Drosophila melanogaster* (DOM), *Danio rerio* (DR), *Homo sapiens* (HS), and *Mus musculus* (MM) are divided into two, with 50% of its data input into the training set and the remaining 50% into the testing set. Then, a five-fold cross-validation technique is employed. Finally, the average performance for the testing data is reported for the proposed system.

DT is a supervised algorithm, and it is generated by using the Iterative Dichotomiser 3 algorithm (ID3) or CART algorithm (Classification algorithm and Regression Tree) (Quinlan, 1986). The DT uses decision nodes to split the dataset into smaller subsets based on information gain (IG) or the Gini index. ID3 uses IG to evaluate how well an attribute splits the training dataset based on its classification objective. IG is the difference between the dataset's entropy before and after splitting depending on the specified attribute values. Let X = $x_1, x_2, x_3, \ldots, x_n$ represent the set of instances, A represent the attribute, and X_v subset of X having A = v. Then, IG is given by

$$IG(X, A) = Ent(X) - \sum_{v \in V(A)} \frac{|X_v|}{|X|} \cdot Ent(X_v), \tag{13}$$

where ENT(X) is the entropy of X and V(A) is the collection of all possible A values. Entropy of X is given by

$$Ent(X) = \sum_{i=1}^{c} -p_i \log_2 p_i, \tag{14}$$

where p_i denotes the probability for current state X.

KNN is a supervised machine learning and non-parametric technique that signifies that it makes no assumptions about the underlying data. The KNN method ensures that the unseen data and

TABLE 7 Impact of features extracted from different co-occurrence matrices for the classification of essential gene sequences of AT, DOM, DR, HS, and MM species.

Classifier	Accuracy	Precision	Recall	F1-score
Effect of first matrix				
K-nearest neighbors	63.37	56.39	63.37	59.20
Decision tree	53.70	54.02	53.70	53.85
Support vector machine	64.38	41.44	64.38	50.42
Effect of second matrix				
K-nearest neighbors	62.05	54.43	62.05	57.54
Decision tree	53.20	53.88	53.20	53.53
Support vector machine	64.38	41.44	64.38	50.42
Effect of third matrix				
K-nearest neighbors	60.58	52.69	60.58	55.66
Decision tree	49.72	51.01	49.72	50.34
Support vector machine	64.38	41.44	64.38	50.42
Effect of fourth matrix				
K-nearest neighbors	62.96	58.32	62.96	59.41
Decision tree	54.33	55.14	54.33	54.72
Support vector machine	64.38	41.44	64.38	50.42
Effect of fifth matrix				
K-nearest neighbors	57.91	49.72	57.91	53.02
Decision tree	47.24	48.14	47.24	47.69
Support vector machine	64.38	41.44	64.38	50.42
Effect of sixth matrix				
K-nearest neighbors	61.49	54.13	61.49	57.14
Decision tree	52.69	54.34	52.69	53.49
Support vector machine	65.35	47.61	65.35	53.36
Effect of seventh matrix				
K-nearest neighbors	58.82	52.94	58.82	55.37
Decision tree	50.44	51.56	50.44	50.99
Support vector machine	64.81	46.81	64.81	53.45
Effect of eighth matrix				
K-nearest neighbors	56.12	50.86	56.12	52.78
Decision tree	49.28	49.86	49.28	49.56
Support vector machine	64.38	41.44	64.38	50.42

existing dataset are comparable and places the unseen data in the most similar class to the unseen data. KNN works by just storing the data during training time. When it sees new data at testing time, it finds k-nearest neighbor to the latest data by using distance measure, i.e., Euclidean distance, and classifies it based on the similarity (Peterson, 2009). The steps of the KNN algorithm are as follows.

1. First, select the value of K, i.e., the closest data points. Any integer may be used as K.
2. Do the following for each data point in the test data set: (i) find the distance between the data point and all samples in the training dataset using one of the following methods: Manhattan, Euclidean, or Hamming distance. In this paper, Euclidean distance measure is used for calculating the distance; (ii) sort samples in the ascending order depending on the distance value; (iii) select the top K samples as the nearest neighbors to the test data point; (iv) next, the test data point will be assigned a class depending on the most common class of these K samples.

The SVM is a supervised machine learning approach for classifying data. The SVM is a well-known technique used in various bioinformatics and computational biology problems, and it needs fewer model parameters to describe the non-linear transition from primary sequence to protein structure region. To minimize the error, the SVM will create the hyperplane repeatedly. The SVM is noted for its quick training, which is necessary for high-throughput database testing (Suthaharan, 2016). Let the dataset be represented by (X_1, y_1), (X_2, y_2), (X_3, y_3), ... , (X_n, y_n). The SVM solves the following equation:

$$\min_{w,b} \|w\|^2 \text{such that} \forall i,\ y_i\left(\langle w, X_i\rangle + b\right) \geq 1, \tag{15}$$

where w and b is the weight and bias of the hyperplane equation $w \cdot X + b = 0$, respectively.

3.3 Evaluation metrics

In this paper, the essential gene classification problem is a multi-class classification problem as we have classified essential genes of five species, i.e., AT, DOM, DR, HS, and MM. For every class in the target, the evaluation matrices (accuracy, precision, recall, and F1-score) were computed. Then, the weighted averaging technique was used to give the final value of evaluation metrics.

$$Accuracy = \frac{\sum_{i=1}^{C} n_i \times \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{\sum_{i=1}^{C} n_i}, \tag{16}$$

$$Precision = \frac{\sum_{i=1}^{C} n_i \times \frac{TP_i}{TP_i + FP_i}}{\sum_{i=1}^{C} n_i} \tag{17}$$

$$Recall = \frac{\sum_{i=1}^{C} n_i \times \frac{TP_i}{TP_i + FN_i}}{\sum_{i=1}^{C} n_i} \tag{18}$$

$$F1 - score = \frac{\sum_{i=1}^{C} n_i \times \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}}{\sum_{i=1}^{C} n_i}, \tag{19}$$

where

$$Precision_i = \frac{TP_i}{TP_i + FP_i}, \tag{20}$$

and

TABLE 8 Impact of existing and proposed features on the classification of essential genes for the AT, DOM, DR, HS, and MM species.

Classifier	Accuracy	Precision	Recall	F1-score
Effect of Shannon entropy features				
K-nearest \neighbors	53.10	46.24	53.10	49.14
Decision tree	48.28	46.96	48.28	47.53
Support vector machine	64.33	41.38	64.33	50.36
Effect of Hurst exponent features				
K-nearest neighbors	53.98	45.63	53.98	49.14
Decision tree	43.57	45.41	43.57	44.45
Support vector machine	64.33	41.38	64.33	50.36
Effect of modified Shannon entropy features				
K-nearest neighbors	54.67	46.20	54.67	49.71
Decision tree	41.76	43.98	41.76	42.80
Support vector machine	64.26	45.64	64.26	50.66
Effect of fractal dimension features				
K-nearest neighbors	58.11	52.19	58.11	52.15
Decision tree	68.35	46.72	68.35	55.51
Support vector machine	68.35	46.72	68.35	55.51
Effect of proposed features				
K-nearest neighbors	64.95	59.49	64.95	61.50
Decision tree	58.31	59.24	58.31	58.70
Support vector machine	66.14	56.57	66.14	54.35

$$Recall_i = \frac{TP_i}{TP_i + FN_i}, \tag{21}$$

where TP_i, TN_i, FP_i, and FN_i are the counts of true positives, true negatives, false positives, and false negatives, respectively, for the i^{th} class. Here, C represents the number of classes in the problem, and n_i indicates the number of samples in the i^{th} class.

3.4 Model framework

The proposed model classified essential genes of five species based on co-occurrence matrices. The proposed model finds the eight different co-occurrence matrices from the DNA sequences. From each co-occurrence matrix, five features, i.e., energy, entropy, homogeneity, contrast, and dissimilarity, were extracted. The existing features, such as HE, FD, SE, and MSE were also computed and then combined with the proposed features for the classification of essential genes. A supervised machine learning algorithm, SVM, was used to evaluate the model. Figure 1 shows essential genes. A supervised machine learning algorithm, SVM was used to evaluate the model. Figure 1 shows the framework of the proposed model.

4 Result and discussion

The proposed essential gene classification model can identify novel essential genes with high recall and precision while only requiring a small number of previously identified essential genes in some species. Such a method could be highly beneficial when investigating essential genes in newly sequenced genomes of other species with few known examples of essential genes. The proposed work has been implemented in the 'Python' environment, while the 'Python' library of machine

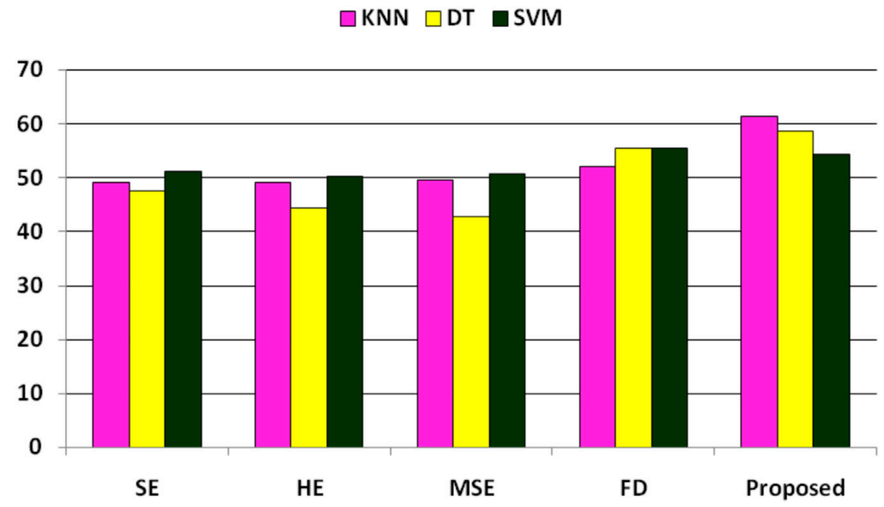


FIGURE 3
Performance (F1-score) comparison of existing features and the proposed features for the classification of essential genes of AT, DOM, DR, HS, and MM species.

TABLE 9 Demonstration of discriminant features among proposed features, Shannon entropy, Hurst exponent, modified Shannon entropy and fractal dimension features.

Feature	Eigen-values	Rank	Feature	Eigen-values	Rank
f_1	13.908	1	f_{23}	0.283	23
f_2	4.434	2	f_{24}	0.257	24
f_3	3.628	3	f_{25}	0.224	25
f_4	2.895	4	f_{26}	0.192	26
f_5	2.505	5	f_{27}	0.152	27
f_6	2.233	6	f_{28}	0.109	28
f_7	1.904	7	f_{29}	0.041	29
f_8	1.602	8	f_{30}	0.032	30
f_9	1.388	9	f_{31}	0.027	32
f_{10}	1.133	10	f_{32}	0.027	31
f_{11}	0.986	11	f_{33}	0.023	33
f_{12}	0.855	12	f_{34}	0.019	34
f_{13}	0.820	13	f_{35}	0.015	35
f_{14}	0.750	14	f_{36}	0.008	36
f_{15}	0.714	15	f_{37}	0.006	37
f_{16}	0.525	16	f_{38}	0.001	43
f_{17}	0.471	17	f_{39}	0.001	44
f_{18}	0.440	18	f_{40}	0.002	42
f_{19}	0.432	19	f_{41}	0.003	41
f_{20}	0.333	20	f_{42}	0.003	40
f_{21}	0.329	21	f_{43}	0.004	39
f_{22}	0.299	22	f_{44}	0.004	38

learning algorithms has been employed for data classification tasks. Python is the best scripting and programming language, is open-source, and has high-level object-oriented programming approaches that deal with mathematical and statistical functions. The method's implementation for the proposed methodology is executed in the Kaggle repository that explores research to data scientists and machine learning engineers as best practitioners in these fields. Here, for Python tools, we have employed NumPy, Pandas, Matplotlib, Sklearn.Preprocessing, Sklearn.Classifiers, Sklearn.Metrics, and some other packages for data analysis and prediction models. The feature vectors extracted from each DNA gene sequence $\mathcal{S}$ undergo KNN, DT, and SVM classifiers. The datasets from AT, DOM, DR, HS, and MM species are given in Table 5. The experimentation of the proposed methodology has been divided into sub-sections.

4.1 Experiment for the proposed features

In this section, experiments with individual features have been performed. Here, from each DNA sequence $\mathcal{S}$, individual feature from each f_I, f_J, f_K, f_L, f_M, f_N, f_O, f_P have been considered, and then classification has been performed. Figure 2 demonstrates the distribution of F1-score performance obtained by DT, KNN, and SVM classifiers with respect to every 40 features computed from co-occurrence matrices of DNA sequence *S*. From this figure, it has been observed that both the KNN and SVM classifiers predict the classification problem better than the DT classifier for most of the features. Moreover, it has also been observed that classifiers have obtained more or less similar performance for most features but better performance due to the 19th, 26th, 27th, 30th, 32nd, and 35th features of the forty-dimensional feature vector *f*. For measuring the impact of individual features such as entropy, homogeneity, energy, contrast, and dissimilarity on the classification of essential genes, the performance has been reported concerning KNN, DT, and SVM classifiers in Table 6. Here, experiments are carried out under the same training–testing protocols, and from each DNA sequence $\mathcal{S}$, the corresponding features are extracted from all co-occurrence matrices. So, each eight-dimensional feature vector is extracted for entropy, homogeneity, energy, contrast, and dissimilarity features.

As shown in Table 6, for every feature, the performance is more or less the same, but for the KNN classifier, the performance is better than that of DT and SVM. Here, F1-score has been considered classification performance as the employed species AT, DOM, DR, HS, and MM have class imbalance problems. Furthermore, the effect of features computed from each co-occurrence matrix in the subsequent experiments has been considered. Here, the 5-dimensional feature vector is extracted from each co-occurrence matrix. The performance due to these feature vectors is reported in Table 7 under the same training–testing protocol. Table 7 shows that there is a more or less a similar effect of co-occurrence matrix features on the essential gene classification. Hence, the features computed from the co-occurrence metrics are helpful and effective. Here, the KNN classifier has better performance.

4.2 Experiment for the existing features

In the further experiment, the performance has been compared with some existing state-of-the-art feature extraction techniques such as SE, MSE, HE, and FD(discussed in Section 2), where these features are extracted accordingly. The performance is obtained concerning KNN, DT, and SVM classifiers. The performance due to these features is reported in Table 8, implying that SE, HE, MSE, and FD features have more or less similar performance. Still, among the classifiers, SVM has obtained better performance. The comparison of these performances and the proposed system has been shown in Figure 3, which shows that the proposed approach has better classified the essential genes of AT, DOM, DR, HS, and MM species under the same training–testing protocol. Here, the difference is in the proposed system, and the forty-dimensional feature vector is considered, while the one-dimensional feature vector is extracted in each existing feature extraction technique. Hence, this work investigates the discriminatory power of co-occurrence matrix features with better performance than the existing state-of-the-art features.

TABLE 10 Demonstration of performance due to combination of features for the classification of essential genes of AT, DOM, DR, HS, and MM species.

Variation	Classifier	Accuracy	Precision	Recall	F1-score	Feature dimension
0.85	K-nearest neighbors	72.01	66.37	72.01	68.67	4
	Decision tree	63.09	63.63	63.09	63.34	
	Support vector machine	74.30	68.77	74.30	67.69	
0.9	K-nearest neighbors	71.52	66.77	71.52	68.94	5
	Decision tree	62.67	63.81	62.67	63.18	
	Support vector machine	75.91	69.57	75.91	70.31	
0.95	K-nearest neighbors	73.82	68.83	73.82	70.80	7
	Decision tree	63.93	64.67	63.93	64.29	
	Support vector machine	76.46	72.63	76.46	71.06	
0.99	K-nearest neighbors	73.96	68.29	73.96	70.66	9
	Decision tree	64.48	65.35	64.48	64.88	
	Support vector machine	76.32	70.56	76.32	**71.42**	

The bold value indicates the highest F1-score.

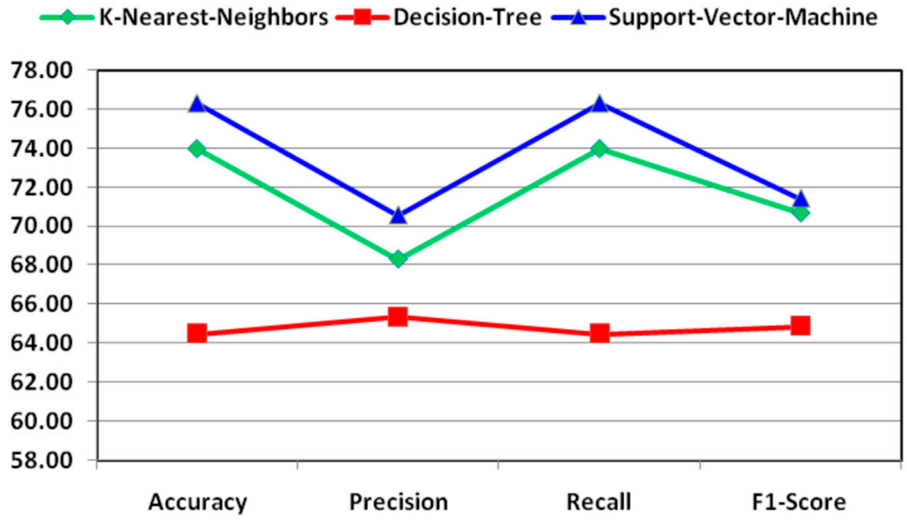


FIGURE 4
Demonstration of final performance for the combination of features for the classification of essential genes of AT, DOM, DR, HS, and MM species.

4.3 Experiment for the combined features

The co-occurrence of nucleotides A, C, T, G in the essential gene derives the distribution of these nucleotides and also their relative position information within the gene $\mathcal{S}$. The existing state-of-the-art techniques of feature extraction (discussed in this work) are key measures in information theory. For example, SE and its modified technique compute the amount of uncertainty and randomness of nucleotides in the gene $\mathcal{S}$. HE measures the relative tendency and characteristic parameters for analyzing its distribution in the essential gene. The FD computes the fractal-like distribution of nucleotides from the indicator matrix calculated from the essential gene $\mathcal{S}$. So, the similarity of patterns of nucleotides computed by the co-occurrence matrices and the information of uncertainty, randomness, relative tendency, and fractal-like distribution information in $\mathcal{S}$ are combined here to obtain more discriminant features for the classification of essential genes of AT, DOM, DR, HS, and MM species. The principal component analysis of dimensionality reduction with variation ratio has been adopted to find the best suitable combination of these features. The performance due to the combination of these features is demonstrated in Table 9.

Table 10 reports the discriminatory power of combined features with respect to various dimensional reduced features concerning

KNN, DT, and SVM classifiers and shows that highest F1-score is 71.42 and it is due to the SVM classifier. As this is class imbalance problem, so F1-score performance has been reported.

For better understanding and visibility, the final performance for the combination of features for the classification of essential genes of AT, DOM, DR, HS, and MM species has been shown in Figure 4.

5 Conclusion

A novel method of feature extraction and analysis for the classification of essential genes of *Arabidopsis thaliana* (AT), *Drosophila melanogaster* (DOM), *Danio rerio* (DR), *Homo sapiens* (HS), and *Mus musculus* (MM) species has been considered in this work. The implementation of the proposed scheme is divided into three segments. In the first segment, novel co-occurrence matrix-based features are extracted from genes that derive the distribution of nucleotides and their relative position from the respective gene. The features from these measures belong to the statistical analysis of the distribution of stationary patterns of nucleotides in the essential genes. In the second segment, some existing state-of-the-art feature computation techniques such as SE, HE, and FD are used as information theory measures that compute uncertainty, randomness, relative tendency, and fractal-like structures in the gene. In the third segment of this work, the features from the proposed methodology and the existing techniques are individually carried out for classification tasks where their F1-score performance has been considered for comparison. These comparisons show the robustness and effectiveness of the proposed methodology. Finally, the features from the proposed scheme and the existing techniques are combined to compute more discriminatory features for classifying essential genes of AT, DOM, DR, HS, and MM species.

Data availability statement

Data used for this study is publicly available at http://www.essentialgene.org/.

Author contributions

RR and SU conceived the method and design. RR, SU, and MK conducted the experiment, and RR, SU, MK, SP, and SM analyzed the results. RR, SU, MK, and SP wrote the manuscript. SM, BB, and HQ reviewed and edited the manuscript. All authors read and approved the final manuscript.

Funding

HQ thanks the United States NSF award 1761839 and 2200138 and a catalyst award from the United States National Academy of Medicine.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1154120/full#supplementary-material

References

Cattani, C. (2010). Fractals and hidden symmetries in dna. *Math. problems Eng.* 2010. 10.1155/2010/507056.

Chen, H., Zhang, Z., Jiang, S., Li, R., Li, W., Zhao, C., et al. (2020). New insights on human essential genes based on integrated analysis and the construction of the hegiap web-based platform. *Briefings Bioinforma.* 21, 1397–1410. doi:10.1093/bib/bbz072

Chen, Y., and Xu, D. (2005). Understanding protein dispensability through machine-learning analysis of high-throughput data. *Bioinformatics* 21, 575–581. doi:10.1093/bioinformatics/bti058

Cullen, L. M., and Arndt, G. M. (2005). Genome-wide screening for gene function using rnai in mammalian cells. *Immunol. cell Biol.* 83, 217–223. doi:10.1111/j.1440-1711.2005.01332.x

Deng, J. (2015). "An integrated machine-learning model to predict prokaryotic essential genes," in *Gene essentiality* (Springer), 137–151.

Dickerson, J. E., Zhu, A., Robertson, D. L., and Hentges, K. E. (2011). Defining the role of essential genes in human disease. *PloS one* 6, e27368. doi:10.1371/journal.pone.0027368

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Véronneau, S., et al. (2002). Functional profiling of the saccharomyces cerevisiae genome. *nature* 418, 387–391. doi:10.1038/nature00935

Gil, R., Silva, F. J., Peretó, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* 68, 518–537. doi:10.1128/MMBR.68.3.518-537.2004

Guo, H.-B., Ghafari, M., Dang, W., and Qin, H. (2021). Protein interaction potential landscapes for yeast replicative aging. *Sci. Rep.* 11, 7143–7154. doi:10.1038/s41598-021-86415-8

Hassan, S. S., Rout, R. K., Sahoo, K. S., Jhanjhi, N., Umer, S., Tabbakh, T. A., et al. (2021). A vicenary analysis of sars-cov-2 genomes. *Cmc-Computers Mater. Continua* 69, 3477–3493. doi:10.32604/cmc.2021.017206

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.* 116, 770–799. doi:10.1061/taceat.0006518

Itaya, M. (1995). An estimation of minimal genome size required for life. *FEBS Lett.* 362, 257–260. doi:10.1016/0014-5793(95)00233-y

Juhas, M., Eberl, L., and Glass, J. I. (2011). Essence of life: Essential genes of minimal genomes. *Trends cell Biol.* 21, 562–568. doi:10.1016/j.tcb.2011.07.005

Juhas, M., Reuß, D. R., Zhu, B., and Commichau, F. M. (2014). Bacillus subtilis and escherichia coli essential genes and minimal cell factories after one decade of genome engineering. *Microbiology* 160, 2341–2351. doi:10.1099/mic.0.079376-0

Juhas, M., Stark, M., von Mering, C., Lumjiaktase, P., Crook, D. W., Valvano, M. A., et al. (2012). High confidence prediction of essential genes in burkholderia cenocepacia. *PloS one* 7, e40064. doi:10.1371/journal.pone.0040064

Khandelwal, M., Kumar Rout, R., Umer, S., Mallik, S., and Li, A. (2022a). Multifactorial feature extraction and site prognosis model for protein methylation data. *Briefings Funct. Genomics* 22, 20–30. doi:10.1093/bfgp/elac034

Khandelwal, M., Rout, R. K., and Umer, S. (2022b). Protein-protein interaction prediction from primary sequences using supervised machine learning algorithm. In 2022 12th International Conference on Cloud Computing, Data Science and Engineering (Confluence) (IEEE), 268–272.

Khandelwal, M., Sheikh, S., Rout, R. K., Umer, S., Mallik, S., and Zhao, Z. (2022c). Unsupervised learning for feature representation using spatial distribution of amino acids in aldehyde dehydrogenase (aldh2) protein sequences. *Mathematics* 10, 2228. doi:10.3390/math10132228

Koonin, E. V. (2000). How many genes can make a cell: The minimal-gene-set concept. *Annu. Rev. genomics Hum. Genet.* 1, 99–116. doi:10.1146/annurev.genom.1.1.99

Kuang, S., Wei, Y., and Wang, L. (2021). Expression-based prediction of human essential genes and candidate lncrnas in cancer cells. *Bioinformatics* 37, 396–403. doi:10.1093/bioinformatics/btaa717

Le, N. Q. K., Do, D. T., Hung, T. N. K., Lam, L. H. T., Huynh, T.-T., and Nguyen, N. T. K. (2020). A computational framework based on ensemble deep neural networks for essential genes identification. *Int. J. Mol. Sci.* 21, 9070. doi:10.3390/ijms21239070

Liu, X., Wang, B.-J., Xu, L., Tang, H.-L., and Xu, G.-Q. (2017). Selection of key sequence-based features for prediction of essential genes in 31 diverse bacterial species. *PLoS One* 12, e0174638. doi:10.1371/journal.pone.0174638

Marques de Castro, G., Hastenreiter, Z., Silva Monteiro, T. A., Martins da Silva, T. T., and Pereira Lobo, F. (2022). Cross-species prediction of essential genes in insects. *Bioinformatics* 38, 1504–1513. doi:10.1093/bioinformatics/btac009

McCutcheon, J. P., and Moran, N. A. (2010). Functional convergence in reduced genomes of bacterial symbionts spanning 200 my of evolution. *Genome Biol. Evol.* 2, 708–718. doi:10.1093/gbe/evq055

Mobegi, F. M., Zomer, A., De Jonge, M. I., and Van Hijum, S. A. (2017). Advances and perspectives in computational prediction of microbial gene essentiality. *Briefings Funct. genomics* 16, 70–79. doi:10.1093/bfgp/elv063

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia* 4, 1883. doi:10.4249/scholarpedia.1883

Qin, H. (2019). Estimating network changes from lifespan measurements using a parsimonious gene network model of cellular aging. *Bmc Bioinforma.* 20, 599–608. doi:10.1186/s12859-019-3177-7

Quinlan, J. R. (1986). Induction of decision trees. *Mach. Learn.* 1, 81–106. doi:10.1007/bf00116251

Rout, R. K., Pal Choudhury, P., Maity, S. P., Daya Sagar, B., and Hassan, S. S. (2018). Fractal and mathematical morphology in intricate comparison between tertiary protein structures. *Comput. Methods Biomechanics Biomed. Eng. Imaging and Vis.* 6, 192–203. doi:10.1080/21681163.2016.1214850

Roemer, T., Jiang, B., Davison, J., Ketela, T., Veillette, K., Breton, A., et al. (2003). Large-scale essential gene identification in candida albicans and applications to antifungal drug discovery. *Mol. Microbiol.* 50, 167–181. doi:10.1046/j.1365-2958.2003.03697.x

Rout, R. K., Ghosh, S., and Choudhury, P. P. (2014). Classification of mer proteins in a quantitative manner. *Int. Comput. Appl. Eng. Sci.* 4, 31–34.

Rout, R. K., Hassan, S. S., Sheikh, S., Umer, S., Sahoo, K. S., and Gandomi, A. H. (2022). Feature-extraction and analysis based on spatial distribution of amino acids for sars-cov-2 protein sequences. *Comput. Biol. Med.* 141, 105024. doi:10.1016/j.compbiomed.2021.105024

Rout, R. K., Hassan, S. S., Sindhwani, S., Pandey, H. M., and Umer, S. (2020). Intelligent classification and analysis of essential genes using quantitative methods. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)* 16, 1–21. doi:10.1145/3343856

Senthamizhan, V., Ravindran, B., and Raman, K. (2021). Netgenes: A database of essential genes predicted using features from interaction networks. *Front. Genet.* 12, 722198. doi:10.3389/fgene.2021.722198

Seringhaus, M., Paccanaro, A., Borneman, A., Snyder, M., and Gerstein, M. (2006). Predicting essential genes in fungal genomes. *Genome Res.* 16, 1126–1135. doi:10.1101/gr.5144106

Suthaharan, S. (2016). "Support vector machine," in *Machine learning models and algorithms for big data classification* (Springer), 207–235.

Umer, S., Dhara, B. C., and Chanda, B. (2016). Texture code matrix-based multi-instance iris recognition. *Pattern Analysis Appl.* 19, 283–295. doi:10.1007/s10044-015-0482-2

Umer, S., Mohanta, P. P., Rout, R. K., and Pandey, H. M. (2021). Machine learning method for cosmetic product recognition: A visual searching approach. *Multimedia Tools Appl.* 80, 34997–35023. doi:10.1007/s11042-020-09079-y

Upadhayay, P. D., Agarwal, R. C., Rout, R. K., and Agrawal, A. P. (2019). Mathematical characterization of membrane protein sequences of homo-sapiens. 2019 9th International Conference on Cloud Computing, Data Science and Engineering (Confluence). IEEE, 382–386.

Veeranagouda, Y., Husain, F., Tenorio, E. L., and Wexler, H. M. (2014). Identification of genes required for the survival of b. fragilis using massive parallel sequencing of a saturated transposon mutant library. *BMC genomics* 15, 429–439. doi:10.1186/1471-2164-15-429

Xu, L., Guo, Z., and Liu, X. (2020). Prediction of essential genes in prokaryote based on artificial neural network. *Genes and genomics* 42, 97–106. doi:10.1007/s13258-019-00884-w

Yuan, Y., Xu, Y., Xu, J., Ball, R. L., and Liang, H. (2012). Predicting the lethal phenotype of the knockout mouse by integrating comprehensive genomic data. *Bioinformatics* 28, 1246–1252. doi:10.1093/bioinformatics/bts120

Zhang, X., Xiao, W., and Xiao, W. (2020). Deephe: Accurately predicting human essential genes based on deep learning. *PLoS Comput. Biol.* 16, e1008229. doi:10.1371/journal.pcbi.1008229

Zurek, W. H. (1989). Algorithmic randomness and physical entropy. *Phys. Rev. A* 40, 4731–4751. doi:10.1103/physreva.40.4731

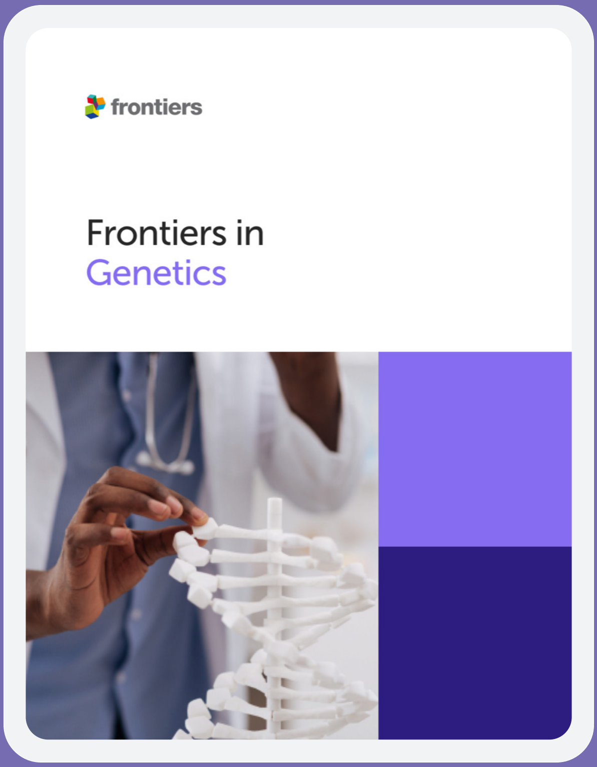
Frontiers in Genetics

Highlights genetic and genomic inquiry relating to all domains of life

The most cited genetics and heredity journal, which advances our understanding of genes from humans to plants and other model organisms. It highlights developments in the function and variability of the genome, and the use of genomic tools.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Genetics

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact